A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean[1]

by

Jim Berger
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #480

September 1977

[1]Research supported by the National Science Foundation under Grants #MCS 76-06627 and MCS 76-06627A2, and the John Simon Guggenheim Memorial Foundation.

Abstract

It is observed that in selecting an alternative to the usual maximum likelihood estimator, δ^0, of a multivariate normal mean, it is important to take into account prior information. Prior information in the form of a prior mean and a prior covariance matrix is considered. A generalized Bayes estimator is developed which is significantly better than δ^0 if this prior information is correct and yet is very robust with respect to misspecification of the prior information. An associated confidence region is also constructed, and is shown to have very attractive size and probability of coverage.

KEY WORDS: Robust generalized Bayes estimators; Multivariate normal mean; Quadratic loss; Risk; Confidence ellipsoids; Size; Probability of coverage.

Table of Contents

Section 1. Introduction

Let $X = (X_1,\ldots,X_p)^t$ have a p-variate normal distribution with mean vector $\theta = (\theta_1,\ldots,\theta_p)^t$ and nonsingular covariance matrix Σ. (Σ will be assumed known until Section 5.) It is desired to estimate θ using an estimator $\delta(X) = (\delta_1(X),\ldots,\delta_p(X))^t$ and under a quadratic loss $L(\delta,\theta) = (\delta-\theta)^tQ(\delta-\theta)$, Q being a positive definite (p×p) matrix. Two common problems giving rise to this setup are (i) estimating a multivariate mean where X is the vector of sample means, and (ii) estimating a vector, θ, of regression coefficients where $X = (B^tB)^{-1}B^tY$ is the least squares estimator and $\Sigma = \sigma^2(B^tB)^{-1}$, B being the design matrix and σ^2 the variance of the errors in the observation Y.

The usual estimator $\delta^o(X) = X$ has been observed to have several deficiencies. These include

1. δ^o is inadmissible if $p \geq 3$. Indeed an estimator δ' can be found with $R(\delta',\theta) < R(\delta^o,\theta) = tr Q\Sigma$ for all θ, where $R(\delta,\theta) = E_\theta L(\delta(X),\theta)$ is the expected loss. This was first noticed by Stein (1955).
2. δ^o doesn't use often existing prior information or relationships among the coordinates, such as when the θ_i are a sample from a superpopulation.
3. When X is the least squares estimator from a regression problem, δ^o is unstable in that (B^tB) is often nearly singular, so that small changes in the observation Y result in very large changes in the estimates of the regression coefficients. (This problem has given rise to the theory of ridge regression, introduced by Hoerl and Kennard (1970).)

In attempting to improve upon δ^o, a number of different approaches have been taken. For the most part, these can be categorized into three areas, according to the nature of the resulting estimator.

The first category consists of approaches resulting in estimators which are linear (i.e. of the form $\delta(X) = CX + \mu$, C a matrix and μ a vector). For example, the Bayesian approach with normal priors and the original form of ridge regression (with a fixed ridge constant) give rise to linear estimators.

The second category of approaches consists of those for which coordinates of θ are set equal to zero, the remainder being estimated in a standard way. For example, preliminary test estimators and typical regression procedures which select the "significant" regression coefficients (effectively setting the others equal to zero) are of this type.

The third category consists of approaches leading to estimators which satisfy

$$\delta(x) = (I - B/(x^tCx))x + o(|x|^{-1}) \quad \text{as } |x| \to \infty \,, \tag{1.1}$$

where B and C are (p×p) matrices, $|x|$ is the Euclidean norm of x, and "o" is the usual little oh notation. For example, minimax, empirical Bayes, Bayes with t-like priors, and stochastic ridge regression approaches all result in estimators of the form (1.1). (By stochastic ridge regression is meant those ridge techniques which choose the ridge constant from the data, using the inverse of some quadratic form in X.)

A number of articles dealing with the above approaches are listed in the references. Unfortunately, the number of articles is by now too large to allow discussion of each contribution specifically, and even too large to reasonably include all in the references. Therefore, only the latest articles and articles specifically referred to are listed. References to earlier works can be found in these articles.

In looking for an alternative to δ^o, only the third category of estimators will be considered. Linear estimators have the well known disadvantage of a lack of robustness with respect to the assumptions under which they are derived. For example, if a Bayesian approach with a normal prior were taken, the resulting linear estimator would have infinite Bayes risk if the true prior were Cauchy. (Even for bounded losses the results could be bad.) In contrast, estimators of the form (1.1) tend to be considerably more robust. Some evidence of this will be presented later. (See also Rubin (1977).)

Estimators from the second category will not be considered for two reasons. First, if indeed estimation is the sole goal, then it has generally been found that discontinuous procedures (such as preliminary test estimators) can be improved upon by smooth shrinkage procedures satisfying (1.1). Of course, there are often compelling reasons (in regression for example) to try for model simplification by setting "nonsignificant" coordinates equal to zero. The goal then is not simply estimation, however, and it seems simplest to approach the problem in two stages - decide first which coordinates are to be set equal to zero, and then use a good estimation procedure on the remaining coordinates. The first stage is outside of the scope of this paper, while for the second stage using a smooth estimator is desirable.

In choosing among estimators of the form (1.1), one is presented with a wide array of principles to go by. The key in choosing among these principles lies in observing the behavior of the estimators - namely, that the estimators perform well (have risk significantly better than δ^o) only in specific regions of the parameter space R^p. Outside these regions they

have risks which are either essentially equivalent to or possibly worse than δ^0. (This is basically due to the fact that δ^0 is minimax, so that no uniformly large improvement in risk is possible.) Since the region of significant improvement differs from estimator to estimator, it seems inescapable that choosing an estimator can be best done by choosing the region of θ over which improvement in risk is desired. In other words, prior knowledge must come into play in effectively choosing an estimator of the form (1.1). (As we shall see this prior knowledge can be quite vague, such as merely believing that the prior distribution of the θ_i is exchangeable.)

Note that the above reasoning is not the usual rationality argument for being Bayesian, but instead a seemingly inevitable conclusion of the particular problem being considered. Indeed if it is felt that there is no prior information whatsoever available, then δ^0 might as well be used, since the "chance" that θ would happen to be in the region of significant improvement of a competing estimator would be negligible. In the remainder of the paper comments will often be phrased in Bayesian terms, not necessarily because a prior distribution on θ is thought to exist, but because it seems necessary to act as if one does exist if a good alternative to δ^0 is to be chosen.

The above considerations also point out the difficulty in meaningfully comparing estimators of the form (1.1) by numerical studies. In numerical studies, the θ at which the estimators are evaluated must be chosen in some fashion, and different estimators will perform best depending on how the θ are chosen. This point was raised by Efron and Morris, Bingham and Larntz, and Thisted in the discussion of Dempster, Schatzoff, and Wermuth (1977).

It is obviously unfeasible to specify prior information for a particular problem, and then choose among all available estimators of the form (1.1) according to which does best for that particular set of prior beliefs. Instead, an estimator should be developed which allows the direct incorporation of prior information in order to adjust its region of significant improvement. This and other desirable properties of an estimator are listed below.

1. δ should readily allow incorporation of prior information.

2. δ should be robust with respect to misspecification of prior information. Equivalently, δ should not have risk $R(\delta,\theta)$ seriously worse than $R(\delta^o,\theta) = \mathrm{tr}(Q\Sigma)$ over a significant region of the parameter space.

3. δ should be expressible in a closed form, relatively simple formula, not only for ease of calculation but also to enable examination for unintuitive or unappealing features.

4. δ should be stable in a ridge sense (providing this is consistent with 1.)

5. δ should be admissible (or nearly so).

6. δ should have the following "empirical Bayes" property. Assume $\Sigma = \sigma^2 I$ and that the θ_i are a random sample from a prior distribution with mean 0 and variance τ^2. Then $\lim_{p\to\infty} |X|^2/p = \sigma^2 + \tau^2$ with probability one. The estimator

$$\delta(X) = (1 - p\sigma^2/|X|^2)X \tag{1.2}$$

is thus very close to the optimal linear Bayes estimator $\delta^L(X) = (1 - \sigma^2/(\sigma^2 + \tau^2))X$, while having a risk uniformly better than δ^o - a very desirable situation. (See Efron and Morris (1973a) for further discussion.)

7. δ should have good associated confidence regions for θ.

The rationale for property 1 has been discussed. Property 2 is also crucial, in that while it is necessary to make use of prior information to significantly improve upon δ^o, we do not want to run the risk of being significantly worse than δ^o if the (often vague) prior information is wrong. Property 3 seems important, partly to make the estimator more attractive to practitioners, but also to make a thorough analysis of the estimator possible. Properties 4, 5, and 6 are all appealing, but perhaps will not be compelling to some statisticians, depending on their philosophical viewpoint. Property 7 is of considerable importance in typical applications of estimation. Section 3 will be devoted to the development and analysis of an interesting set of confidence regions.

In attempting to verify the above properties for a proposed estimator, numbers 2, 5, and 7 cause the most difficulty. In checking #2, Berger (1976b) can be useful, though numerical studies are probably necessary. The only certain method of ensuring that #5 is satisfied is to develop δ as an admissible generalized Bayes estimator. (Brown (1971) shows that an estimator must be generalized Bayes to be admissible.) Trying to verify that an estimator is "nearly" admissible is difficult. A useful negative result is given in Berger and Srinivasan (1977), namely that estimators satisfying (1.1) are approximations to generalized Bayes estimators (up to a $o(|x|^{-1})$term) if and only if $B = k \not\in C$ for some constant k.

Estimators so far proposed do not fully satisfy the above list of properties. The only estimators of the form (1.1) which allow the incorporation of prior information are empirical Bayes estimators (see for example Efron and Morris (1973a) and Rolph (1976)) and Bayes estimators arising

from flat priors (see for example Leonard (1976)). Unfortunately the estimators which have been developed using these approaches cannot be written in closed form (except for a few special cases of Q, $\hat{L}$, and prior information), making a meaningful analysis of them very difficult. In Section 2 a reasonable generalized Bayes estimator is developed which does satisfy the above 7 properties.

Before proceeding, a word is in order as to what type of prior input is envisaged. Recall that the real goal is to decide what region of the parameter space is of greatest importance. A relatively simple approach would be to specify an ellipsoid of interest. This ellipsoid could be written as $\{\theta: (\theta-\mu)^t A^{-1}(\theta-\mu) \leq p\}$. Alternatively it seems plausible to assume the availability of a prior mean vector, μ, for θ, and also of a variance (or covariance) matrix A which reflects the believed accuracy of the guess, μ. In either case the prior input is conveniently summarized by μ and A. Only rarely will additional prior knowledge (such as knowledge of the functional form of the prior) be available. Hence it is desired to construct an estimator which can make use of μ and A, but which requires no further knowledge of the prior in order to be better than δ^o. The estimator should also be robust in the sense that if μ and A do not reflect the true value of θ (or the true prior of θ for Bayesians), the estimator should not be significantly worse than δ^o.

Further discussion of the prior input is given in Section 4, where it is shown how to incorporate into the above framework such things as a belief in exchangeability of the prior, or a belief that certain linear restrictions on θ hold. Until then it will also be assumed that $\mu = 0$, a simple translation which saves considerably on notation.

In the following sections the notation $\det(B)$, $tr(B)$, and $ch_{max}(B)$ will be used to denote the determinant, trace, and maximum characteristic root of

a matrix B. Also, E will be used to denote expectation, with subscripts denoting parameter values and superscripts denoting random variables with respect to which the expectation is to be taken. When obvious, subscripts and superscripts will be deleted.

Section 2. The Generalized Bayes Estimator

2.1. Development of the Estimator

Let C be a (p×p) symmetric matrix such that $(C-\Sigma)$ is positive semi-definite. Define $B(\lambda) = \lambda^{-1}C-\Sigma$, for $\lambda > 0$. For $n > 0$, consider the generalized prior density

$$g_n(\theta) = \int_0^1 [\det\{B(\lambda)\}]^{-1/2}\exp\{-\theta^t B(\lambda)^{-1}\theta/2\}\lambda^{(n-1-p/2)}d\lambda. \tag{2.1}$$

Note that the conditional density of θ given λ is normal with mean 0 and covariance matrix $B(\lambda)$, while λ has the (generalized) density $\lambda^{(n-1-p/2)}$ on (0,1). This prior is a generalization of one considered in Berger (1976a), and for $\Sigma = C = I$ was first introduced by Strawderman (1971). (Judge and Bock (1977) give a good discussion of these special cases.)

Several aspects of g_n are interesting to observe. First, it can be shown that asymptotically (for large $|\theta|$) g_n behaves like $k(\theta^t C^{-1}\theta)^n$ for some constant k. Thus larger n correspond to "sharper tails" for the prior. It can also be checked that g_n has finite mass for $n > p/2$.

For certain C, n, and p, $g_n(\theta)$ can be calculated explicitely. For example, if $C = c\Sigma$ $(c \geq 1)$, $p = 4$, and $n = (p-2)/2 = 1$, then

$$g_n(\theta) = k(1-\exp\{-\theta^t\Sigma^{-1}\theta/[2(c-1)]\})/\theta^t\Sigma^{-1}\theta.$$

The actual form of g_n is not of great importance, however, being as we don't really think that g_n is the true prior. It is just being used as a tool to develop an estimator which we hope will exhibit desirable behavior.

The generalized Bayes estimator of θ, with respect to g_n, is given by

$$\delta^n(X) = \frac{\int \theta \exp\{-(X-\theta)^t \Sigma^{-1}(X-\theta)/2\} g_n(\theta) d\theta}{\int \exp\{-(X-\theta)^t \Sigma^{-1}(X-\theta)/2\} g_n(\theta) d\theta}.$$

It is straightforward to check that $g_n(\theta)$ has finite mass over any compact neighborhood of zero. This, along with the fact that $g_n(\theta)$ is bounded outside a neighborhood of zero, allows interchanging the order of integration in the numerator above to get

$$\int \theta \exp\{-(X-\theta)^t \Sigma^{-1}(X-\theta)/2\} g_n(\theta) d\theta$$

$$= \int_1^0 \int \theta \exp\{-[(X-\theta)^t \Sigma^{-1}(X-\theta) + \theta^t B(\lambda)^{-1}\theta]/2\} d\theta \, [\det\{B(\lambda)\}]^{-1/2} \lambda^{(n-1-p/2)} d\lambda .$$

Completing squares and integrating out over θ in the last expression results in the equivalent formula

$$\int_0^1 [(\Sigma^{-1} + B(\lambda)^{-1})^{-1}\Sigma^{-1}X] \exp\{-X^t[\Sigma^{-1} - \Sigma^{-1}(\Sigma^{-1} + B(\lambda)^{-1})^{-1}\Sigma^{-1}]X/2\}$$

$$\times [\det(\Sigma^{-1} + B(\lambda)^{-1})]^{-1/2} [\det B(\lambda)]^{-1/2} \lambda^{(n-1-p/2)} d\lambda.$$

Using the matrix identities

$$[\Sigma^{-1} + B(\lambda)^{-1}]B(\lambda) = \Sigma^{-1}B(\lambda) + I = \Sigma^{-1}C/\lambda,$$

$$[\Sigma^{-1} + B(\lambda)^{-1}]^{-1} = \Sigma - \Sigma[\Sigma + B(\lambda)]^{-1}\Sigma \tag{2.2}$$

$$= \Sigma - \Sigma(C/\lambda)^{-1}\Sigma = \Sigma - \lambda\Sigma C^{-1}\Sigma,$$

it can be concluded that

$$\int \theta \exp\{-(X-\theta)^t \Sigma^{-1}(X-\theta)/2\} g_n(\theta)\, d\theta$$

$$= \int_0^1 (I - \lambda \Sigma C^{-1}) X \exp\{-\lambda X^t C^{-1} X/2\} [\det(\Sigma^{-1} C)]^{-1/2} \lambda^{n-1} d\lambda .$$

A similar calculation verifies that

$$\int_0^1 \exp\{-(X-\theta)^t \Sigma^{-1}(X-\theta)/2\} g_n(\theta)\, d\theta$$

$$= \int_0^1 \exp\{-\lambda X^t C^{-1} X/2\} [\det(\Sigma^{-1} C)]^{-1/2} \lambda^{n-1} d\lambda . \tag{2.3}$$

Hence, defining $||X||^2 = X^t C^{-1} X$ and

$$r_n(v) = \frac{v \int_0^1 \lambda^n \exp\{-\lambda v/2\} d\lambda}{\int_0^1 \lambda^{(n-1)} \exp\{-\lambda v/2\} d\lambda}, \tag{2.4}$$

it follows that

$$\delta^n(X) = (I - \frac{r_n(||X||^2)\Sigma C^{-1}}{||X||^2}) X. \tag{2.5}$$

An integration by parts in the numerator of (2.4) establishes that

$$r_n(v) = 2n(1 - [n \int_0^1 \lambda^{n-1} \exp\{-(\lambda-1)v/2\}]^{-1}). \tag{2.6}$$

Integration by parts also shows that

$$[n \int_0^1 \lambda^{n-1} \exp\{-(\lambda-1)v/2\} d\lambda]^{-1} = [\sum_{i=0}^{\infty} \frac{\Gamma(n+1)(v/2)^i}{\Gamma(n+1+i)}]^{-1} \tag{2.7}$$

$$= \begin{cases} \dfrac{(v/2)^n}{n![\exp\{v/2\} - \sum_{i=0}^{n-1} (v/2)^i/i!]} & \text{if } n \text{ is an integer} \\ \dfrac{(v/2)^n}{\Gamma(n)[\exp\{v/2\}\mathrm{erf}\{(v/2)^{1/2}\} - \sum_{i=0}^{(n-3/2)} (v/2)^{(i+1/2)}/\Gamma(i+3/2)} & \text{if } n - 1/2 \text{ is an integer,} \end{cases}$$

where $\mathrm{erf}(z) = (2/\sqrt{\pi}) \int_0^z \exp\{-t^2\}dt$. The last expressions in (2.7) are obviously particularly convenient for calculation. The following lemma gives several useful properties of r_n.

Lemma 2.1.1. If $n > 0$, then

(i) $0 \leq r_n(v) < 2n$.

(ii) $r_n(v)$ is increasing in v.

(iii) $\lim_{v\to\infty} r_n(v) = 2n$.

(iv) $\lim_{v\to 0} [r_n(v)/\{nv/(n+1)\}] = 1$.

(v) $\lim_{n\to\infty} r_n(v) = v$.

(vi) $\lim_{n\to\infty} [r_n(2nc)/(2n\{\min(1,c)\})] = 1$.

(vii) $r_n(v)/v$ is decreasing in v.

(viii) $\lim_{v\to\infty} [r_n'(v)/\{\exp(-v/2)(v/2)^n/\Gamma(n)\}] = 1$, where $r_n'(v) = \frac{d}{dv} r_n(v)$.

(ix) $\lim_{v\to\infty} v^m r_n'(v) = 0$ for all $m > 0$.

(x) $r_n(v)/v \leq n/(n+1)$.

(xi) $\lim_{v\to\infty} v^m(r_n(v) - 2n) = 0$ for all $m > 0$.

Proof. Parts (i), (ii), and (iii) follow immediately from (2.6) and the first expression in (2.7). Parts (iv) and (v) follow from the first expression in (2.7), after noticing that the first two terms of the summation are dominant as $v \to 0$ or $n \to \infty$.

To prove part (vi), the first expression in (2.7) will again be used. For fixed i, the ith term of the summation satisfies

$$\lim_{n\to\infty} [(nc)^i \Gamma(n+1)/\Gamma(n+1+i)] = c^i.$$

Hence

$$\lim_{n\to\infty} \Big[\sum_{i=0}^{\infty} \frac{(nc)^i \Gamma(n+1)}{\Gamma(n+1+i)}\Big]^{-1} = \Big[\sum_{i=0}^{\infty} c_i\Big]^{-1} = \begin{cases} 0 & \text{if } c \geq 1 \\ (1-c) & \text{if } 0 < c < 1. \end{cases}$$

The result follows.

To prove part (vii), consider λ as a random variable with density

$$\exp\{-\lambda v/2\}\lambda^{(n-1)} I_{(0,1)}(\lambda) / \int_0^1 \exp\{-\lambda v/2\}\lambda^{(n-1)} d\lambda,$$

where $I_{(0,1)}(\lambda)$ is the indicator function on (0,1). It is easy to check that the above density has decreasing monotone likelihood ratio in v, and hence that the expected value of λ must be decreasing in v. But from (2.4) it is clear that the expected value of λ is simply $r_n(v)/v$, and the conclusion follows.

To prove part (viii), observe that a calculation using (2.6) gives

$$r_n'(v) = \frac{\exp\{-v/2\} \int_0^1 \lambda^{(n-1)}(1-\lambda)\exp\{-\lambda v/2\} d\lambda}{\Big[\int_0^1 \lambda^{(n-1)} \exp\{-\lambda v/2\} d\lambda\Big]^2}. \tag{2.8}$$

But

$$\int_0^1 \lambda^m \exp\{-\lambda v/2\}d\lambda = (v/2)^{-(m+1)} \int_0^{v/2} \lambda^m \exp\{-\lambda\}d\lambda$$

$$= (v/2)^{-(m+1)}(\Gamma(m+1) - o(v^{-1})).$$

Using this in (2.8) gives the desired result.

Part (ix) follows immediately from part (viii). Part (x) follows from parts (iv) and (vii). Part (xi) follows from (2.6) and the first expression in (2.7). ||

The first question which arises is how should n and C be chosen? The choice of n that is recommended is n = (p-2)/2. The estimator δ^n can then be easily calculated using (2.5) and (2.7), and the resulting estimator will be seen to have many nice properties. Note that by Lemma 2.1.1 (iii), $\lim_{v\to\infty} r_n(v) = 2n = (p-2)$ for this choice of n. When $C = \Sigma = I$, the estimator $\delta^{(p-2)/2}(X)$ is thus similar to the original Stein estimator $\delta(X) = (1 - [p-2]/|X|^2)X$. Further justification for this choice of n will be seen later.

As a guide in choosing C, note that the covariance matrix of $g_n(\theta)$ (for $n > \frac{p}{2} + 1$) is given by ($(n-p/2)$ is the normalizing constant for g_n)

$$\Lambda = (n - \tfrac{p}{2}) \int \theta\,\theta^t g_n(\theta)d\theta$$

$$= (n - \tfrac{p}{2}) \int_0^1 \int \theta\,\theta^t [\det\{B(\lambda)\}]^{-1/2} \exp\{-\theta^t B(\lambda)^{-1}\theta/2\}d\theta\lambda^{(n-1-p/2)}d\lambda$$

$$= (n - \tfrac{p}{2}) \int_0^1 B(\lambda)\lambda^{(n-1-p/2)}d\lambda$$

$$= (n - \tfrac{p}{2}) \int_0^1 [\tfrac{C}{\lambda} - \Sigma]\lambda^{(n-1-p/2)}d\lambda$$

$$= \frac{(n-p/2)}{(n-1-p/2)}\, C - \Sigma.$$

Hence if A is felt to be the true covariance matrix of the prior (see Section 1), then it is reasonable to equate A and Λ. The implied choice of C is

$$C = \frac{(n-1-p/2)}{(n-p/2)} (\ddagger + A).$$

While we will mainly be interested in $n = (p-2)/2$ (for which g_n has infinite mass and hence no covariance matrix), the above considerations suggest choosing $C = \rho(\ddagger + A)$ for some constant $\rho \geq ch_{max}\{\ddagger(\ddagger+A)^{-1}\}$. (Note that the condition on ρ is necessary to ensure that $C \geq \ddagger$.)

An alternative viewpoint which suggests choosing C as above is to consider the situation where the prior is known to have mean zero and covariance matrix A. In such a definite setting one would probably be quite happy to use the best linear estimator (in terms of Bayes risk.) This best linear estimator is easily calculated to be

$$\delta(X) = (I - \ddagger(\ddagger + A)^{-1})X.$$

Choosing $C = \rho(\ddagger + A)$ results in

$$\delta^n(X) = (I - \frac{r_n(||X||^2)\ddagger(\ddagger + A)^{-1})}{X^t(\ddagger + A)^{-1}X} X,$$

where $||X||^2 = X^t(\ddagger + A)^{-1}X/\rho$. This estimator "corrects" X in the direction $\ddagger(\ddagger + A)^{-1}X$ exactly as does the best linear estimator, but controls the amount of correction in a way which is quite reasonable. To see this, note that if A is the "correct" prior covariance, then $\lim_{p\to\infty} X^t(\ddagger + A)^{-1}X/p = 1$ with probability one. Hence $||X||^2 \cong p/\rho$ for large p. ("$\cong$" denotes approximate equality.) By Lemma 2.1.1 (vi) it follows that for large p,

$$r_{(p-2)/2}(||X||^2) \tilde{=} p(\min\{1, 1/\rho\}).$$

Thus if A is correct, p is large, and

$$\text{ch}_{\max}\{\sharp(\sharp + A)^{-1}\} \leq \rho \leq 1, \tag{2.9}$$

then

$$\delta^{(p-2)/2}(||X||^2) \tilde{=} (I - \sharp(\sharp + A)^{-1})X$$

as would be desired. If on the other hand A is wrong, or θ is not in the region expected, then $[X^t(\sharp + A)^{-1}X]$ will tend to be much larger than $r_{(p-2)/2}(||X||^2)$, and $\delta^{(p-2)/2}$ will correct $\delta^o(X) = X$ very little.

The above considerations are not meant to prove anything, but merely to indicate why the suggested estimator is reasonable. Note in particular that choosing $2n \tilde{=} p$ (as is $n = (p-2)/2$), was necessary to obtain the desired convergence to the best linear estimator for large p.

A decision must also be made as to what value of ρ (satisfying (2.9)) to use. Note that ρ affects δ^n only through $r_n(X^t(\sharp + A)^{-1}X/\rho)$. It is clear from Lemma 2.1.1 (ii) that r_n is decreasing in ρ, so that larger ρ result in more conservative estimators (in that they are closer to $\delta^o(X) = X$). There are no apparent theoretical guidelines to help in the choice of ρ (for $n = (p-2)/2$), so a variety of numerical studies were performed (some of which will be seen later). Roughly, it was found that $\rho = .6$ gave the best overall performance in terms of Bayes risks when A is large. For smaller A, choosing $\rho = 2\ \text{ch}_{\max}\{\sharp(\sharp + A)^{-1}\}$ worked well. Keeping in mind the restrictions in (2.9), the following estimator is thus recommended:

$$(2.10) \qquad \delta^*(X) = \left(I - \frac{r^*(X^t(\Sigma + A)^{-1}X/\rho^*)\Sigma(\Sigma + A)^{-1}}{X^t(\Sigma + A)^{-1}X}\right)X,$$

where $r^* = r_{(p-2)/2}$ and

$$\rho^* = \min\{1, \max[2\lambda, .6]\} = \begin{cases} 1 & \text{if } \lambda \geq .5 \\ 2\lambda & \text{if } .3 \leq \lambda \leq .5, \\ .6 & \text{if } \lambda \leq .3 \end{cases}$$

where $\lambda = ch_{max}\{\Sigma(\Sigma + A)^{-1}\}$.

2.2 Evaluation of δ^*.

This estimator δ^* will now be examined carefully to see if it satisfies properties 1 through 7 given in Section 1. Some of what follows pertains to the whole class of estimators δ^n, while some refers specifically to δ^*. Which is being discussed will clearly be indicated.

Property 1. δ^* readily allows the incorporation of prior knowledge as was the main goal. The question arises as to whether the incorporation of the prior knowledge, A, leads to a significant improvement for estimators of this form (assuming the prior knowledge is approximately correct). To investigate this question p-variate normal priors, $\xi(\theta)$, with mean 0 and covariance matrix τB were considered. (Note that these priors are not really close to the priors $g_n(\theta)$ in terms of tail behavior. Therefore, we are not loading the dice in favor of the estimator δ^*.) The Bayes risks $r(\delta,\xi) = \int R(\delta,\theta)\xi(\theta)d\theta$ of three estimators, $\delta^{\tau B}$, δ^B, and δ^I were compared. $\delta^{\tau B}$ is δ^* with the "correct" choice $A = \tau B$. δ^B is δ^* with $A = B$, meaning the wrong scale factor is being used. δ^I is δ^* with $A = I$, so that an entirely wrong covariance matrix is being used. Typical of the numerical results

obtained are those given in the first three rows of Table 1. The calculations there are for $p = 6$, $Q = \Sigma = I$, and B diagonal with diagonal elements $\{.1, .5, 1, 3, 6, 16\}$. (Note that this is a wide spread of variances (for $\tau = 1$ anyway), in that some coordinates have comparatively small sample variance, some have comparatively small prior variance, and some are in between.) For varying τ, the Bayes risks of $\delta^{\tau B}$, δ^{B}, and δ^{I} are given in Table 1. $\delta^{\tau B}$ is clearly best, while δ^{B} is significantly better than δ^{I}. Thus it appears that the incorporation of prior knowledge in δ^* is quite worthwhile, though it need not be absolutely correct in order to achieve significant gains. (The Bayes risk of the usual estimator is $r(\delta^{o},\xi) = 6$.)

Table 1. Bayes Risks

	τ								
	.25	.50	.75	1.0	2.0	5.0	10.0	25.0	50.0
$\delta^{\tau B}$	3.44	3.88	4.14	4.32	4.72	5.18	5.46	5.69	5.82
δ^{B}	3.88	4.05	4.20	4.32	4.74	5.30	5.61	5.84	5.92
δ^{I}	4.04	4.52	4.82	5.02	5.43	5.71	5.85	5.93	5.97
$\delta_L^{\tau B}$	2.16	2.82	3.21	3.47	4.08	4.77	5.12	5.58	5.76
δ_L^{B}	2.78	3.01	3.24	3.47	4.39	7.15	11.75	25.56	48.56
δ_L^{I}	3.16	4.83	6.49	8.15	14.80	34.80	68.00	167.8	334.0

Property 2. δ^* is quite robust with respect to misspecification of prior information, and has a risk $R(\delta^*,\theta)$ which compares quite favorably with $R(\delta^{o},\theta)$.

The robustness is indicated by Table 1. δ^{B} and δ^{I} use (at say $\tau = 50$) drastically wrong prior information, and yet still have better Bayes risks than δ^*. Indeed it can be shown that $\lim_{\tau\to\infty} r(\delta^*,\xi) = 6$ no matter what fixed A

is used in δ^*. This compares quite favorably with the corresponding situation when linear Bayes estimators are used. The estimators $\delta_L^{\tau B}$, δ_L^B, and δ_L^I in Table 1 are the linear estimators defined by

$$\delta_L^A(X) = (I - \ddagger(\ddagger + A)^{-1})X.$$

Thus $\delta_L^{\tau B}$ is the optimum linear and indeed optimum Bayes estimator for the situation of Table 1. δ_L^B and δ_L^I correspond to misspecified prior information. The risks given in Table 1 show the nonrobustness of the linear estimators compared to δ^*. The case for estimators such as δ^* would be even more telling if priors with flat tails were used. (We are looking at linear estimators on their home ground so to speak.)

Studies of Bayes risks alone tend to put estimators such as δ^* in a very flattering light. To discover the seamier side of such estimators, it is important to look at the regular risk $R(\delta^*,\theta)$ in comparison with $R(\delta^o,\theta)$. Since δ^* generally pulls $\delta^o(X) = X$ closer to zero, it can be expected that $R(\delta^*,\theta) < R(\delta^o,\theta)$ for θ in a neighborhood of zero. It also usually turns out to be true that $R(\delta^*,\theta) < R(\delta^o,\theta)$ for θ in certain directions of the parameter space. The reverse inequality can hold in other directions. Of usefulness in analyzing this behavior are the results of Berger (1976b). (See also Brown (1974)). Using Theorem 1, Lemma 1, and Lemma 2 of Berger (1976b), together with Lemma 2.1.1 (i, iii, and ix) of this paper, it can be shown that

$$\Delta(\theta) = R(\delta^n,\theta) - R(\delta^o,\theta) = \frac{-4n}{\theta^t C^{-1}\theta}\left\{\mathrm{tr}(\ddagger Q \ddagger C^{-1}) - \frac{(2+n)\theta^t C^{-1}\ddagger Q \ddagger C^{-1}\theta}{\theta^t C^{-1}\theta}\right\} + o(|\theta|^{-2}). \tag{2.11}$$

Note immediately that $\Delta(\theta) \to 0$ as $|\theta| \to \infty$. Furthermore, if $|\theta|$ is large enough, it follows that $\Delta(\theta) < 0$ if

$$(2.12) \qquad \frac{(2+n)\theta^t C^{-1} \ddagger Q \ddagger C^{-1}\theta}{\theta^t C^{-1}\theta} < \operatorname{tr}(\ddagger Q \ddagger C^{-1}).$$

This can be written in a more enlightening fashion by letting $\Gamma = C^{-1/2} \ddagger Q \ddagger C^{-1/2}$, letting $\{\alpha_i\}$ denote the eigenvalues of Γ, letting $\{v_i\}$ denote an associated set of orthonormal eigenvectors, and expressing θ as

$$(2.13) \qquad \theta = (\theta^t C^{-1}\theta)^{1/2} \sum_{i=1}^{p} e_i C^{1/2} v_i, \text{ where } \sum_{i=1}^{p} e_i^2 = 1.$$

Noting also that $\operatorname{tr}(\ddagger Q \ddagger C^{-1}) = \operatorname{tr}(\Gamma) = \sum_{i=1}^{p} \alpha_i$, it is clear that (2.12) can be rewritten

$$(2.14) \qquad (2+n) \sum_{i=1}^{p} \alpha_i e_i^2 < \sum_{i=1}^{p} \alpha_i.$$

From (2.14) can be determined the directions in which $\Delta(\theta) < 0$ for large $|\theta|$. Usually, $R(\delta^n,\theta)$ will be less than $R(\delta^o,\theta)$ for all θ in these directions. Note in particular that if (2.14) holds for all e_i, or equivalently that

$$(2+n)\operatorname{ch}_{\max}(\ddagger Q \ddagger C^{-1}) = (2+n)\max\{\alpha_i\} < \sum_{i=1}^{p} \alpha_i = \operatorname{tr}(\ddagger Q \ddagger C^{-1}),$$

then $\Delta(\theta) < 0$ for large $|\theta|$. Indeed using Theorem 1 of Berger (1976c), the following stronger result can be obtained:

Theorem 2.2.1. If $(2+n)ch_{max}(\Sigma Q \Sigma C^{-1}) \leq tr(\Sigma Q \Sigma C^{-1})$, then $R(\delta^n, \theta) \leq R(\delta^o, \theta)$ for all θ. (δ^n is hence minimax.)

Proof: Since $n > 0$, the condition of the theorem clearly ensures that $p \geq 3$. Assumptions (i) and (ii) of Theorem 1 of Berger (1976c) follow immediately. Assumptions (iii) and (iv) of Berger (1976c) can be verified by a simple calculation of ∇r_n (the gradient of r_n), together with Lemma 2.1.1 (ii, iii, and ix). Assumption (v) of Berger (1976c) is satisfied by the condition given in the theorem. The conclusion follows.||

Corollary 2.2.2. δ^* is minimax if $p \geq 3$ and

$$(2.15) \qquad (p+2)ch_{max}\{\Sigma Q \Sigma(\Sigma + A)^{-1}\} \leq 2\, tr\{\Sigma Q \Sigma(\Sigma + A)^{-1}\}.$$

This is in particular true if

(a.) $Q = c_1 I$, $\Sigma = c_2 I$, $A = c_3 I$, $c_1 > 0$, $c_2 > 0$, $c_3 \geq 0$;

(b.) $Q = c(I + \Sigma^{-1}A)\Sigma^{-1}$, $c > 0$; or

(c.) $A = c\Sigma Q \Sigma - \Sigma$, $c \geq ch_{max}(\Sigma^{-1}Q^{-1})$.

Proof: Immediate.||

Thus whenever (2.15) holds, δ^* will have risk smaller than $R(\delta^o, \theta)$ for all θ, a very nice situation. Note, in particular, that this will be true for the symmetric situation described in (a.) of the Corollary. (The result of Corollary 2.2.1 was obtained in the particular case $A = [ch_{max}(\Sigma^{-1}Q^{-1})\Sigma Q \Sigma - \Sigma]$ in Berger (1976a), and in the case $Q = \Sigma = I$ and $A = 0$ in Strawderman (1971).)

Unfortunately, for nonsymmetric problems (2.15) will not typically be satisfied. It is instructive to examine the risk function $R(\delta^*, \theta)$ in such a situation by numerical methods. The situation considered was $p = 6$, $Q = I$, $A = I$, and Σ diagonal with diagonal elements $\{.1, 1, 1, 1, 1, 10\}$.. This is a

case of considerable nonsymmetry where δ^* can be expected to be worse than δ^o in certain directions of the parameter space. Indeed, calculating Γ shows that $\alpha_1 = 1/110$, $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1/2$, and $\alpha_6 = 100/11$. The $\{v_i\}$ are just the unit vectors on the corresponding axes.Thus (2.14) becomes

$$4[e_1^2/110 + (e_2^2+e_3^2+e_4^2+e_5^2)/2 + 100e_6^2/11] < 11.1, \tag{2.16}$$

and for $\{e_i\}$ satisfying this equation we would expect that $\Delta(\theta) < 0$ for θ in the directions (see (2.13))

$$(1.05e_1,\ 1.41e_2,\ 1.41e_3,\ 1.41e_4,\ 1.41e_5,\ 3.32e_6)^t.$$

The risk funetion $R(\delta^*,\theta)$ was numerically computed along the six coordinate axes and along the line $\theta = |\theta|(1,1,1,1,1,1)^t/6^{1/2}$. From (2.16), we would expect that $\Delta(\theta) < 0$ along all these lines except the θ_6 axis. The numerical results in Figure 1 bear this out. ($R(\delta^o,\theta) = 14.1$ is the constant line on the graph.)

The risk of δ^* along the θ_6 axis appears to be seriously worse that that of δ^o, but recall it is being assumed that A = I. Thus the prior belief is roughly that θ_6 has mean 0 and variance one. If indeed θ_6 turns out to be 10 standard deviations away from zero, some penalty must be expected. Note, at least, that the penalty is bounded. For comparison purposes, the risk $R(\delta_L,\theta)$ along the θ_6 axis of the optimum linear Bayes estimator is also given in Figure 1. The comparative robustness of δ^* is clear.

Figure 1.

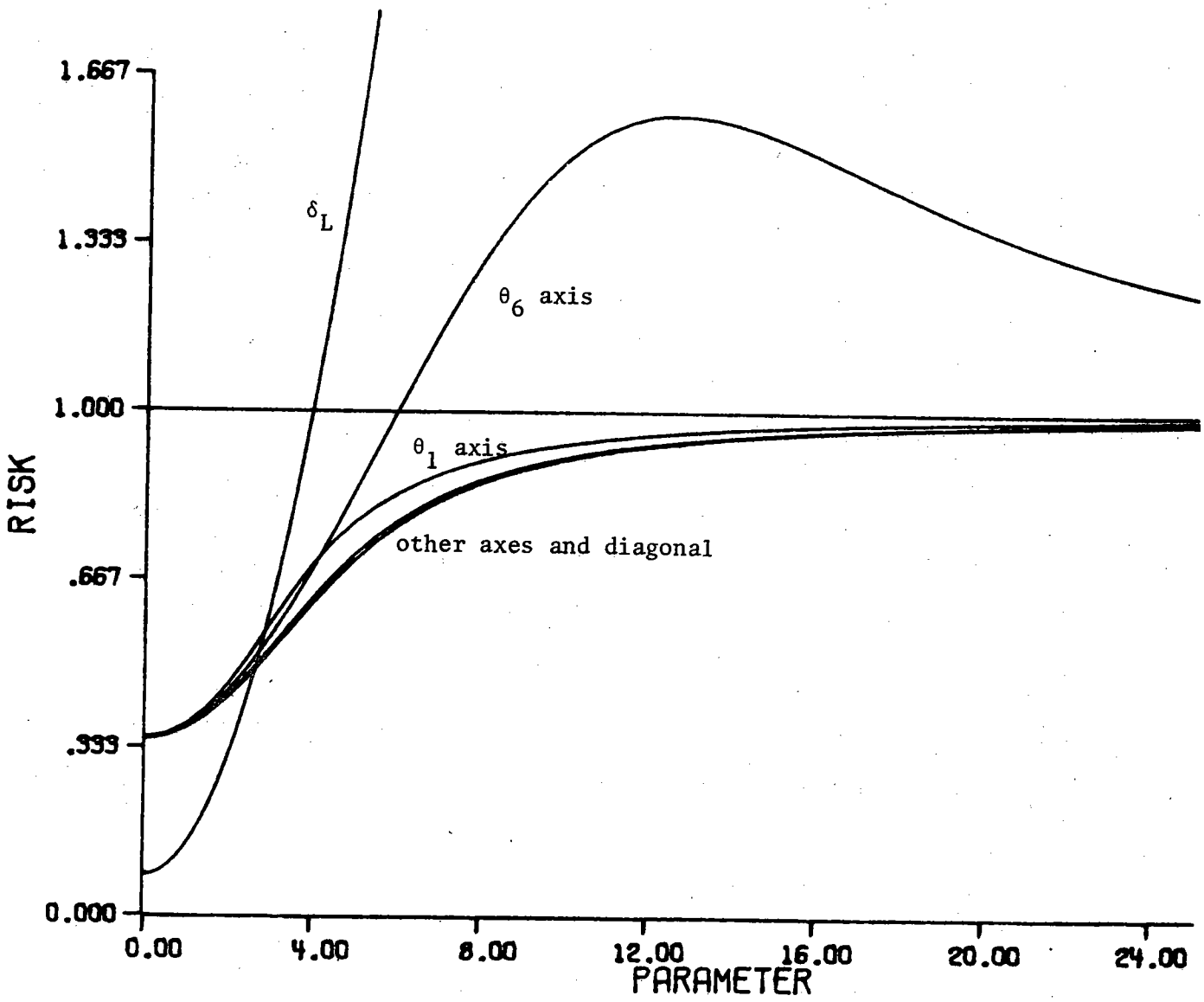


Figure 1 makes it graphically clear that in using δ^* minimaxity will often be sacrificed. It seems reasonable, however, to give up minimaxity in unimportant areas of the parameter space in order to achieve sizeable improvement elsewhere. Minimax estimators do not appear to be able to achieve the sizeable gains in (Bayes) risk offered by δ^* in nonsymmetric problems.

The difficulty with minimax estimators can be seen by examination of the typical minimax estimator (Hudson (1974) and Berger (1976a))

$$\delta^M(X) = (I - \frac{(p-2)Q^{-1}\Sigma^{-1}}{X^t\Sigma^{-1}Q^{-1}\Sigma^{-1}X})X.$$

Clearly coordinates with high variance get pulled in (towards zero) proportionally less than coordinates with low variance. This is contrary to the intuitive idea that inaccurate X_i should be corrected more than accurate X_i. This problem seems to be common to all minimax estimators, with the result being that little improvement is obtained in nonsymmetric problems. (Hudson (1974), Thisted (1976), Morris (1977), and Casella (1977) also discuss this problem.)

δ^* has another advantage over minimax estimators which is of a more practical nature. This is that the loss matrix Q need not be known in order to calculate δ^*. (On the other hand, Q plays a crucial role in all minimax estimators.) In applications it is usually much easier to obtain prior information (like u and A) from a client, than it is to obtain Q. (People will readily guess where θ is, but are reluctant to say which coordinates are more important than others.) This point was also made by Morris (1977).

Property 3. δ^* is clearly relatively easy to calculate, use, and analyze.

Property 4. Stochastic ridge estimators make no formal allowance for prior information, but they are similar to δ^n with the choices $C = [ch_{max}(\Sigma)]I$ and $n = p/2$. Hence estimators δ^n can be found with about the same "stability" as stochastic ridge estimators. (See Casella (1977) for some definitions of stability). The prior input into an estimator seems far more important than its stability, however, so no attempt was made in choosing δ^* to force it to be stable.

Property 5. As in Berger (1976a), the results of Brown (1971) (in particular Theorem 6.4.2) can be used to show that δ^n is admissible if $n \geq (p-2)/2$, and inadmissible if $n < (p-2)/2$. Thus δ^* is admissible.

As indicated previously, the flatter the tails of a prior are, the more robust the generalized Bayes estimator derived from that prior tends to be. Since, for $g_n(\theta)$, smaller n correspond to flatter tails, it appears that δ^* is about as robust as possible (in terms of choice of n), while preserving admissibility. This was another reason for choosing $n = (p-2)/2$.

Property 6. The discussion leading to the choice of δ^* in Section 2.1 showed that δ^* has a crude empirical Bayes property - namely that if A is chosen correctly and p is large, then δ^* is approximately the optimum linear Bayes estimator. For the symmetric empirical Bayes situation discussed in Section 1, the following stronger empirical Bayes property can be obtained.

Assume $\Sigma = \sigma^2 I$, $A = cI$, and the θ_i are a sample from a prior distribution with mean zero and variance τ^2. Note that for δ^*,

$$\lim_{p\to\infty} \frac{||X||^2}{p} = \frac{|X|^2}{p\rho^*(\sigma^2+c)} = \frac{(\sigma^2+\tau^2)}{\rho(\sigma^2+c)} \quad \text{with probability one.}$$

Lemma 2.1.1 (vi) can be used to conclude that

$$\lim_{p\to\infty} r^*(||X||^2) = p \min\{1, \frac{\sigma^2+\tau^2}{\rho^*(\sigma^2+c)}\} \quad \text{with probability one.}$$

Hence $\delta^*(X)$ behaves like (1.2) as desired, providing $(\sigma^2+\tau^2)/[\rho^*(\sigma^2+c)] \geq 1$. This last equality would hold for all τ^2 if ρ^* had been chosen to be $\sigma^2/(\sigma^2+c)$ $(= ch_{max}\{\Sigma(\Sigma+A^{-1})\})$. The choice given in (2.10) was deemed more desirable, however. Note in any case that since $\rho^* \leq 1$, $(\sigma^2+\tau^2)/[\rho^*(\sigma^2+c)] \geq 1$ if $\tau^2 > c$.

Property 7. The next section deals with confidence regions for θ based on δ^*.

Section 3. Confidence Regions for θ.

While there has been a great deal of research on multivariate estimation of θ, there has been comparatively little on the development of improved confidence regions for θ. The theoretical works of Brown (1966) and Joshi (1967) established that the usual confidence region could be improved upon, but did not provide explicit improved confidence regions. By the usual confidence region is meant

$$C^o(X) = \{\theta: \ (X-\theta)^t \Sigma^{-1} (X-\theta) \leq k(\alpha)\},$$

where $k(\alpha)$ is the $100(1-\alpha)$th percentile of the chi square distribution with p degrees of freedom. Stein (1962) and (1974) suggests certain confidence regions for large p (based on heuristic considerations), but leaves open the question of what to do for small or moderate p. Faith (1976) in the symmetric situation ($\Sigma = A = I$) develops Bayesian confidence regions using priors similar to $g_n(\theta)$, and gives convincing numerical and theoretical arguments to support their superiority over C^o. Unfortunately, his confidence regions are difficult to work with, having a complicated shape arising from their Bayesian derivation.

Morris (1977) suggests an appealing way to proceed in a Bayesian fashion, with a resulting confidence region which is fairly simple. In the symmetric situation he considers the prior $g_n(\theta)$ with $n = (p-2)/2$ and $C = I$, and calculates the posterior mean, $\delta^n(X)$, and posterior covariance matrix, $\Sigma_n(X)$. He uses the diagonal elements of $\Sigma_n(X)$ to construct confidence intervals for the θ_i, centered at $\delta_i^n(X)$. The resulting confidence region is simple and

yet hopefully retains the advantage of a robust Bayesian approach. We will differ somewhat from Morris by considering confidence ellipsoids based on the entire $\Sigma_n(X)$, and of course dealing with the nonsymmetric situation.

3.1. Development of the Confidence Region

The first step is the calculation of $\Sigma_n(X)$, the covariance matrix of the posterior distribution of θ given X (for the prior $g_n(\theta)$). Clearly, $\Sigma_n(X)$ is given by

$$(3.1) \qquad \Sigma_n(X) = \frac{\int [\theta-\delta^n(X)][\theta-\delta^n(X)]^t \exp\{-(X-\theta)^t\Sigma^{-1}(X-\theta)/2\} g_n(\theta)d\theta}{\int \exp\{-(X-\theta)^t\Sigma^{-1}(X-\theta)/2\} g_n(\theta)d\theta}.$$

Using (2.1), completing squares, and interchanging orders of integration as in Section 2, gives that the numerator of (3.1) is

$$(3.2) \qquad \int_0^1 \exp\{-X^t[\Sigma^{-1}-\Sigma^{-1}(\Sigma^{-1}+B(\lambda)^{-1})^{-1}\Sigma^{-1}]X/2\}[\det B(\lambda)]^{-1/2}\lambda^{(n-1-p/2)}$$

$$\times \int_{R^p} [\theta\theta^t-\delta^n(X)\delta^n(X)^t]\exp\{-(\theta-z)^t(\Sigma^{-1}+B(\lambda)^{-1})(\theta-z)/2\}d\theta d\lambda,$$

where $z = (\Sigma^{-1}+B(\lambda)^{-1})^{-1}\Sigma^{-1}X$. Replacing $[\theta\theta^t]$ by $[(\theta-z)(\theta-z)^t+\theta z^t+z\theta^t-zz^t]$ and integrating over θ, the inside integral in (3.2) is equal to

$$[\det(\Sigma^{-1}+B(\lambda)^{-1})]^{-1/2}[(\Sigma^{-1}+B(\lambda))^{-1}+zz^t-\delta^n(X)\delta^n(X)^t].$$

Using this along with the identities in (2.2) and the definitions of $\delta^n(X)$ and z, the expression (3.2) can be calculated to be

$$(3.3) \qquad \int_0^1 \exp\{-\lambda||X||^2/2\}\lambda^{n-1}[\det(\Sigma^{-1}C)]^{-1/2}[\Sigma-\lambda\Sigma\, C^{-1}\Sigma$$

$$+\left(\frac{r_n(||x||^2)}{||x||^2} - \lambda\right)(\sharp C^{-1}xx^t + xx^t C^{-1}\sharp) + \left(\lambda^2 - \frac{r_n^2(||x||^2)}{||x||^4}\right)\sharp C^{-1}xx^t C^{-1}\sharp]d\lambda.$$

Using (2.3), (2.4), (3.1), (3.3), and defining

$$t_n(v) = \frac{v^2 \int_0^1 \exp\{-\lambda v/2\}\lambda^{n+1} d\lambda}{\int_0^1 \exp\{-\lambda v/2\}\lambda^{n-1} d\lambda}, \tag{3.4}$$

it follows that

$$\sharp_n(X) = \sharp - \frac{r_n(||x||^2)}{||x||^2}\sharp C^{-1}\sharp + \frac{[t_n(||x||^2) - r_n^2(||x||^2)]}{||x||^4}\sharp C^{-1}xx^t C^{-1}\sharp. \tag{3.5}$$

Integration by parts in the numerator of (3.4) establishes that

$$t_n(v) = 4n(n+1)\left\{1 - \frac{1+v/\{2(n+1)\}}{n\int_0^1 \exp\{-(\lambda-1)v/2\}\lambda^{n-1} d\lambda}\right\}. \tag{3.6}$$

The integral in the above expression can be evaluated using (2.7). For calculational purposes, it is probably easier to observe from (2.6) that

$$t_n(v) = 2(n+1)r_n(v) - v(2n - r_n(v)) = r_n(v)[2(n+1)+v] - 2nv. \tag{3.7}$$

The following properties of t_n will be needed.

Lemma 3.1.1. If $n > 0$, then

(i) $0 < t_n(v) < 4n(n+1)$.

(ii) $\lim_{v\to\infty} t_n(v) = 4n(n+1)$.

(iii) $\lim_{v\to 0} [t_n(v)/\{nv^2/(n+2)\}] = 1$.

(iv) $t_n(v) - r_n^2(v) = 2r_n(v) - 2vr_n'(v)$.

(v) $0 < t_n(v) - r_n^2(v) < 2r_n(v)$.

Proof: Parts (i), (ii), and (iii) follow from (3.6) and (2.7) exactly as did the corresponding results in Lemma 2.1.1.||

To prove part (iv), note that differentiating in (2.6) gives

$$r_n'(v) = \frac{r_n(v)}{v} - \frac{t_n(v)}{2v} + \frac{r_n^2(v)}{2v} .$$

Rearranging terms gives the desired result. The upper bound in part (v) follows immediately from (iv) and Lemma 2.1.1 (ii). To establish the lower bound, note that by Lemma 2.1.1 (vii), $\log[r_n(v)/v]$ is decreasing in v. Hence

$$0 > \frac{d}{dv} [\log r_n(v) - \log v] = \frac{r_n'(v)}{r_n(v)} - \frac{1}{v} ,$$

or $r_n(v) - vr_n'(v) > 0$. Part (iv) then completes the argument.||

The following lemma will be needed later on, and provides an interesting bound on $\ddagger_n(X)$. For two (p×p) matrices A and B, let $A \leq B$ mean that (B - A) is positive semidefinite.

Lemma 3.1.2.

$$\ddagger - \frac{n}{(n+1)} \ddagger C^{-1} \ddagger \leq \ddagger_n(X) \leq \ddagger + \frac{n}{(n+1)} \ddagger C^{-1} \ddagger .$$

Proof: The lower bound follows from (3.5), using Lemma 3.1.1 (v) and Lemma 2.1.1 (x). The upper bound follows from Lemma 3.1.1 (v), Lemma 2.1.1 (x), and the fact that $||X||^{-2} C^{-1/2} X X^t C^{-1/2} \leq I$.||

The lower bound in Lemma 3.1.2 is sharp in that $\ddagger_n(0) = \ddagger - \frac{n}{n+1} \ddagger C^{-1} \ddagger$. (This follows from Lemma 3.1.1 (iii) and Lemma 2.1.1 (iv).) The upper bound

is not sharp in that the rank one matrix $C^{-1/2}XX^tC^{-1/2}$ was bounded by the rank p matrix $||X||^2I$.

The confidence regions that will be considered are the ellipsoids

$$C^n(X) = \{\theta \in R^p: \ [\theta-\delta^n(X)]^t \Sigma_n(X)^{-1} [\theta-\delta^n(X)] \leq k(\alpha)\}, \tag{3.8}$$

where $k(\alpha)$ is the $100(1-\alpha)$th percentile of the chi square distribution with p degrees of freedom. Note that these are not the true Bayesian confidence sets for the priors g_n, but are only approximations based on the posterior means and covariances. They do have a familiar shape, however, and are quite easy to work with. In the calculation of $\Sigma_n(X)^{-1}$, the following lemma is useful.

Lemma 3.1.3. If Y is a $(p\times 1)$ vector and B a $(p\times p)$ matrix, then

$$(I + YY^tB)^{-1} = (I - [1 + Y^tBY]^{-1}YY^tB).$$

Proof: Calculation.||

For convenience, define

$$u = u(||X||^2) = \frac{r_n(||X||^2)}{||X||^2}, \quad w = w(||X||^2) = \frac{t_n(||X||^2)-r_n^2(||X||^2)}{||X||^4}. \tag{3.9}$$

Letting $B = (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}$ and $Y = \Sigma C^{-1}X$, Lemma 3.1.3 can be applied to (3.5) to give

$$\begin{aligned}\Sigma_n(X)^{-1} &= (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}(I + w\Sigma C^{-1}XX^tC^{-1}\Sigma[\Sigma - u\Sigma C^{-1}\Sigma]^{-1})^{-1} \\ &= B(I - [1 + wY^tBY]^{-1}wYY^tB).\end{aligned} \tag{3.10}$$

Thus the calculational problem is reduced to finding $B = (\Sigma - u\Sigma C^{-1}\Sigma)^{-1}$. If, in particular, $\Sigma = I$, $C = \rho I$, then

$$(3.11) \qquad \Sigma_n(X)^{-1} = (1-u/\rho)^{-1}(I-wXX^t/[\rho^2-\rho u+w|X|^2]).$$

The particular choice of n and C that is recommended is $n = (p-2)/2$ and $C = \rho^*(\Sigma+A)$ (ρ^* defined in (2.10)), so that the resulting confidence region is centered at δ^*. Let $C^*(X)$, $\Sigma^*(X)$, and t^* denote $C^n(X)$, $\Sigma_n(X)$, and t_n for these choices of n and C. Thus

$$(3.12) \quad \Sigma^*(X) = \Sigma - \frac{r^*(||X||^2)}{||X||^2}\Sigma(\Sigma+A)^{-1}\Sigma + \frac{[t^*(||X||^2)-r^*(||X||^2)^2]}{||X||^4}\Sigma(\Sigma+A)^{-1}XX^t(\Sigma+A)^{-1}\Sigma,$$

where $||X||^2 = X^t(\Sigma+A)^{-1}X/\rho^*$, and

$$(3.13) \qquad C^*(X) = \{\theta: \quad [\theta-\delta^*(X)]^t\Sigma^*(X)^{-1}[\theta-\delta^*(X)] \leq k(\alpha)\}.$$

It is interesting to consider certain intuitive explanations for the terms of $\Sigma^*(X)$. Note first that in the standard Bayesian model where θ has a multivariate normal distribution with mean vector zero and covariance matrix A, the posterior covariance matrix is

$$(3.14) \qquad (\Sigma^{-1}+A^{-1})^{-1} = \Sigma - \Sigma(\Sigma+A)^{-1}\Sigma.$$

In Section 2.1 it was shown that if A is the correct prior covariance matrix and p is large, then $r^*(||X||^2)/(\rho^*||X||^2) \cong 1$. Hence the first two terms of $\Sigma^*(X)$ behave like (3.14) when A is correct and p is large. On the other hand, if the A used is incorrect, then $r^*(||X||^2)/(\rho^*||X||^2)$ will usually be small and $\Sigma^*(X)$ will behave more like Σ. Note that the last term of $\Sigma^*(X)$ is relatively insignificant in large p situations since it is a rank one matrix.

Another appealing facet of the large p behavior of $C^*(X)$ is that for the symmetric situation ($\Sigma = I$, $A = \tau I$) $C^*(X)$ is similar to the confidence

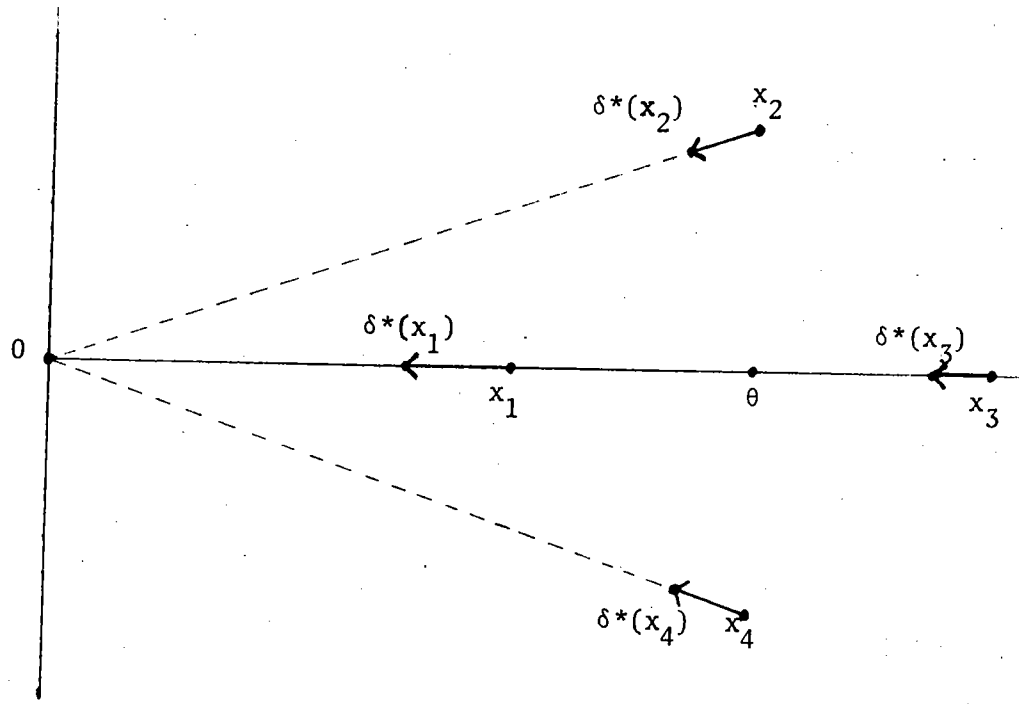
region suggested by Stein (1962). Indeed when $||X||^2 \geq p$ (the likely situation for large p), then $r^*(||X||^2) \tilde{=} p$, so (ignoring the rank one third term)

$$C^*(X) \tilde{=} \{\theta: \ |\theta-(1 - \frac{p}{|X|^2})X|^2 \leq (1 - \frac{p}{|X|^2})k(\alpha)\},$$

which is the confidence region suggested by Stein up to first order terms.

The third term of $\phi^*(X)$ seems rather strange at first sight. It has a very reasonable intuitive explanation, however. Note that the characteristic vector corresponding to the nonzero characteristic root of the third term of $\phi^*(X)$ is $z = \phi(\phi+A)^{-1}X$. Hence in the direction of z, the contribution of the third term is positive. (The confidence ellipsoid is widened.) In directions perpendicular to z the third term is zero and the confidence ellipsoid is narrowed. Note, on the other hand, that δ^* (at which ϕ^* is centered) performs relatively badly when it "corrects" X along the same line that contains θ. (Correcting only along a line results in essentially a one dimensional problem.) δ^* achieves its gains when correcting those X for which the direction of correction is close to perpendicular to $(X-\theta)$. This phenomenon is exhibited in Figure 2, where θ is shown with four symmetrically placed possible X values. Assume the simple estimator $\delta^*(X) = (1 - r^*(|x|^2)/|x|^2)x$ is being used, so the x values will be shrunk towards zero. Clearly the effect of δ^* upon x_1 and x_3 (the x's corrected along the line containing θ) is harmful, in that the average distance of $\delta^*(x_1)$ and $\delta^*(x_3)$ from θ is larger than the average distance of x_1 and x_3 from θ. On the other hand, δ^* moves x_2 and x_4 closer to θ. (This type of picture was shown to me by Lawrence Brown.)

Figure 2.



Since δ^* corrects X in the direction z, our confidence region should reflect the harmful effect δ^* would have upon θ lying in that direction by being widened in that direction. This is precisely how $C^*(X)$ behaves.

Morris (1977) bases his confidence regions only upon the first two terms of $\ddagger^*(X)$ and the diagonal elements of the third term. The above argument indicates this may be undesirable.

We now proceed with a more rigorous analysis of the properties of $C^*(X)$. The two common criteria used in evaluating confidence regions are size and probability of coverage. Size will be considered first. (Many of the mathematical results which follow will be stated for general $\ddagger_n(X)$.)

3.2. Size of $C^n(X)$

There are a number of reasonable measures of the size of an ellipsoid. Virtually all are functions of the lengths of the semiaxes of the ellipsoid. For $C^n(X)$, the lengths of the semiaxes are the characteristic roots of $\hat{\Sigma}_n(X)^{1/2}$.

Actually, it is perhaps more appropriate to be concerned with the roots of $[Q\hat{\Sigma}_n(X)]^{1/2}$, in order to take into account the relative importance of the various coordinates as reflected by Q. This is natural as can be seen by transforming the problem by $Q^{1/2}$ (i.e. define $Y = Q^{1/2}X$, $\eta = Q^{1/2}\theta$, etc.). In the transformed problem the loss is sum of squares error loss so that all coordinates are of equal importance. It is easy to check that the posterior covariance matrix (given Y) in the transformed problem is

$$Q^{1/2}\hat{\Sigma}_n(Q^{-1/2}Y)Q^{1/2} = Q^{1/2}\hat{\Sigma}_n(X)Q^{1/2},$$

and hence it is natural to look at the characteristic roots of the square roots of this matrix, or equivalently the roots of $[Q\hat{\Sigma}_n(X)]^{1/2}$. For those who prefer to consider size of the original $\hat{\Sigma}_n(X)$, merely set $Q = I$ in the results below.

The following three measures of size of $C^n(X)$ will be considered:

1. $\det[(Q\hat{\Sigma}_n(X)]^{1/2} = (\det Q)^{1/2}(\det\hat{\Sigma}_n(X))^{1/2}$, which up to a multiplicative dimensional constant is the volume of the transformed confidence ellipsoid. Clearly it suffices to consider only $[\det\hat{\Sigma}_n(x)]^{1/2}$ since Q occurs only in a multiplicative constant which will be the same for all transformed ellipsoids. Hence comparisons of volumes will be unaffected by Q.

2. $\mathrm{tr}[Q\Sigma_n(X)]^{1/2}$ which is the sum of the semiaxes of the transformed confidence ellipsoid.

3. $\mathrm{tr}[Q\Sigma_n(X)]$, which is the sum of the squares of the semiaxes of the transformed confidence ellipsoid. This measure of size is of additional interest since it is also the posterior expected loss of δ^n.

The results in this section will be concerned with comparing the size of $C^n(X)$ (and $C^*(X)$) to the size of $C^o(X)$, the usual confidence region. Note that for $C^o(X)$, the three measures of size that will be discussed are $\det(\Sigma^{1/2})$, $\mathrm{tr}(Q\Sigma)^{1/2}$, and $\mathrm{tr}(Q\Sigma)$ respectively.

The first result gives a condition on X under which $\Sigma_n(X) < \Sigma$, and hence $C^n(X)$ has smaller size than $C^o(X)$ under any reasonable measure of size. The notation in (3.9) will be used extensively from here on.

Theorem 3.2.1. If $u(||X||^2) > w(||X||^2)$, then $\Sigma_n(X) < \Sigma$.

Proof: This follows immediately from (3.5), noting that $\Sigma C^{-1}XX^tC^{-1}\Sigma \leq ||X||^2 \Sigma C^{-1}\Sigma$. ||

To investigate the first measure of size, the following two lemmas are needed.

Lemma 3.2.2.

$$\det\Sigma_n(X) = [\det\Sigma][\det(I-u(||X||^2)C^{-1}\Sigma)][1+w(||X||^2)X^t(C\Sigma^{-1}C-u(||X||^2)C)^{-1}X].$$

Proof: Clearly

$$\det\Sigma_n(X) = [\det\Sigma][\det(I-uC^{-1}\Sigma+wC^{-1}XX^tC^{-1}\Sigma)] \tag{3.15}$$

$$= [\det\Sigma][\det(I-uC^{-1}\Sigma)][\det(I+wC^{-1}XX^tC^{-1}\Sigma\{I-uC^{-1}\Sigma\}^{-1})].$$

Note that $C^{-1}XX^tB$ has rank one for any nonsingular (p×p) matrix B, and has characteristic roots 0 (with multiplicity (p-1)) and $(X^tBC^{-1}X)$. ($C^{-1}X$ is the characteristic vector of the nonzero root.) The characteristic roots of $[I+wC^{-1}XX^tB]$ are hence 1 (with multiplicity (p-1)) and $(1+wX^tBC^{-1}X)$. It follows that

$$\det(I+wC^{-1}XX^tC^{-1}\ddagger\{I-uC^{-1}\ddagger\}^{-1}) = 1+wX^tC^{-1}\ddagger\{I-uC^{-1}\ddagger\}^{-1}C^{-1}X$$

$$= 1+wX^t(C\ddagger^{-1}C-uC)^{-1}X.$$

Together with (3.15) this gives the desired result.||

Lemma 3.2.3. Assume that $a_i \geq 0$ and $b_i \geq 0$ $(i=1,\ldots,p)$, and that $p \geq 2$, $\sum_{i=1}^{p} b_i = 1$, and $\sum_{i=1}^{p} a_i \geq 2 \max_{1\leq i\leq p}\{a_i\}$. Then

$$(3.16)\qquad \prod_{i=1}^{p}(1+ya_i[2b_i-1]) \leq 1 \text{ for all } y \in [0, (\max_i\{a_i(1-2b_i)\})^{-1}].$$

Proof: Without loss of generality assume that a_1 is the largest a_i. If $b_i \leq 1/2$ for all i, the conclusion is obvious. Hence assume $b_j > 1/2$ for some j. Note then that $b_i < 1/2$ for all $i \neq j$. Examining (3.16), it is clear that the worst case to consider is j = 1(since a_1 is the largest a_i). Thus assume $b_1 > 1/2$.

Since $2a_1 \leq \sum_{i=1}^{p} a_i$ (or $a_1 \leq \sum_{i=2}^{p} a_i$), it is clear that

$$(3.17)\qquad \prod_{i=1}^{p}(1+ya_i[2b_i-1]) \leq (1+y\{\sum_{i=2}^{p} a_i\}[2b_1-1])[\prod_{i=2}^{p}(1+ya_i[2b_i-1])].$$

Denoting the right hand side above by $\varphi(y)$, a calculation gives

$$\frac{d}{dy}\varphi(y) = \{\sum_{i=2}^{p} a_i\}[2b_1-1][\prod_{i=2}^{p}(1+ya_i[2b_i-1])]$$

$$+ [\prod_{i=2}^{p}(1+ya_i[2b_i-1])]\sum_{j=2}^{p}\{\frac{a_j[2b_j-1](1+y\{\sum_{i=2}^{p} a_i\}[2b_1-1])}{1+ya_j[2b_j-1]}\}$$

$$= [\prod_{i=2}^{p}(1+ya_i[2b_i-1])]\sum_{j=2}^{p}\{\frac{a_j[2(b_1+b_j-1)+y(2b_1-1)(2b_j-1)(a_j+\sum_{i=2}^{p} a_i)}{1+ya_j[2b_j-1]}\}.$$

Since $(2b_1-1) > 0$, $(2b_j-1) < 0$, $a_i \geq 0$, $(b_1+b_j-1) \leq 0$, and $(1+ya_j[2b_j-1]) \geq 0$ (due to the domain of y), it is clear that $\frac{d}{dy}\varphi(y) \leq 0$. Hence $\varphi(y)$ is maximized at $y = 0$, which together with (3.17) establishes the result.||

The following theorem gives conditions under which the volume of $C^n(X)$ is less than the volume of $C^o(X)$.

Theorem 3.2.4. $[\det\ddagger_n(X)]^{1/2} \leq [\det\ddagger]^{1/2}$ for all X, if and only if $\mathrm{tr}(C^{-1}\ddagger) \geq 2\mathrm{ch}_{\max}(C^{-1}\ddagger)$.

Proof: Using Lemma 3.2.2, it is clear that showing that $[\det\ddagger_n(X)]^{1/2} \leq [\det\ddagger]^{1/2}$ is equivalent to showing that

$$H = [\det(I-u(||X||^2)C^{-1}\ddagger)][1+w(||X||^2)X^t(C\ddagger^{-1}C-u(||X||^2)C)^{-1}X] \leq 1. \tag{3.18}$$

For convenience, let T be orthogonal such that $T^tC^{-1/2}\ddagger C^{-1/2}T = D$ is diagonal with diagonal elements $\{d_1,\ldots,d_p\}$, d_1 being the largest. Note that the condition $\mathrm{tr}(C^{-1}\ddagger) \geq 2\mathrm{ch}_{\max}(C^{-1}\ddagger)$ is simply

$$\sum_{i=1}^{p} d_i \geq 2d_1. \tag{3.19}$$

Also define $z = T^tC^{-1/2}X$, so that $||X||^2 = X^tC^{-1}X = |z|^2$. Then H can be rewritten

$$(3.20)\qquad H = \Big[\prod_{i=1}^{p}(1-u(|z|^2)d_i)\Big]\big[1+w(|z|^2)z^t(D^{-1}-u(|z|^2)I)^{-1}z\big]$$

$$= \Big[\prod_{i=1}^{p}(1-u(|z|^2)d_i)\Big]\Big[1+w(|z|^2)\sum_{i=1}^{p}\{z_i^2 d_i/(1-u(|z|^2)d_i)\}\Big].$$

To prove the "only if" part of the theorem, choose $z = |z|(1,0,\ldots,0)^t$. Then

$$(3.21)\qquad H = \Big[\prod_{i=1}^{p}(1-u(|z|^2)d_i)\Big]\big[1+w(|z|^2)|z|^2 d_1/(1-u(|z|^2)d_1)\big]$$

$$= \Big[\prod_{i=2}^{p}(1-u(|z|^2)d_i)\Big]\big[1+d_1\{w(|z|^2)|z|^2-u(|z|^2)\}\big].$$

Letting $|z| \to \infty$ and using Lemma 2.1.1 ((iii) and (ix)) and Lemma 3.1.1 (iv), it is clear that

$$u(|z|^2) = r_n(|z|^2)/|z|^2 = 2n/|z|^2 + o(|z|^{-2}),$$

$$w(|z|^2)|z|^2-u(|z|^2) = \frac{2r_n(|z|^2)-2|z|^2 r_n'(|z|^2)}{|z|^2} - \frac{r_n(|z|^2)}{|z|^2}$$

$$= 2n/|z|^2 + o(|z|^{-2}).$$

Hence from (3.21)

$$H = \Big[1-\frac{2n}{|z|^2}\Big(\sum_{i=2}^{p} d_i\Big)+o(|z|^{-2})\Big]\Big[1+\frac{2nd_1}{|z|^2}+o(|z|^{-2})\Big]$$

$$= 1+\frac{2n}{|z|^2}\{d_1-\sum_{i=2}^{p} d_i\} + o(|z|^{-2}).$$

Thus if (3.19) is violated, then $H > 1$ for large enough $|z|$ (and z in the given direction). This proves the "only if" part of the theorem.

To prove the "if" part, observe from (3.20) that

$$H \leq \Big[\prod_{i=1}^{p}(1-ud_i)\Big]\Big[\prod_{i=1}^{p}(1+wz_i^2 d_i/\{1-ud_i\})\Big] \tag{3.22}$$

$$= \prod_{i=1}^{p}(1+d_i[wz_i^2-u])$$

$$\leq \prod_{i=1}^{p}(1+ud_i[\frac{2z_i^2}{|z|^2}-1]),$$

the last step following from Lemma 3.1.1 (v). Letting $y = u$, $a_i = d_i$, $b_i = z_i^2/|z|^2$, and applying Lemma 3.2.3, gives that if (3.19) is satisfied then $H \leq 1$, completing the proof.||

Corollary 3.2.5. If $C = \rho\sharp(\rho \geq 1)$ and $p \geq 2$, then $[\det\sharp_n(X)]^{1/2} \leq [\det\sharp]^{1/2}$ for all X.

Proof: $\mathrm{tr}(C^{-1}\sharp) = p\tau \geq 2\tau = 2\mathrm{ch}_{\max}(C^{-1}\sharp)$, so Theorem 3.2.4 gives the desired result.

Corollary 3.2.6. $C^*(X)$ has smaller volume than $C^o(X)$ for all X, if and only if

$$\mathrm{tr}(I+A\sharp^{-1})^{-1} \geq 2\mathrm{ch}_{\max}(I+A\sharp^{-1})^{-1}.$$

Proof: Obvious from Theorem 3.2.4, noting that $\sharp(\sharp+A)^{-1} = (I+A\sharp^{-1})^{-1}$.||

Note in particular that for the symmetric problem where $\sharp$ and A are multiples of the identity, then $C_n(X)$ and $C^*(X)$ have smaller volume than $C^o(X)$ for $p \geq 2$.

The question arises as to how significant an improvement in volume is obtainable using $C^*(X)$ instead of $C^o(X)$. Using Lemma 3.2.2 it is an easy matter to calculate

$$V^*(X) = \frac{\text{volume of } C^*(X)}{\text{volume of } C^o(X)} = \frac{[\det \Sigma^*(X)]^{1/2}}{[\det \Sigma]^{1/2}} .$$

Typical of the results obtained are those given in Tables 2 and 3 below. Table 2 considers the symmetric situation $\Sigma = I$ and $A = 2I$ (so $C = \rho^*(\Sigma+A) = 2I$) for $p = 6$ and $p = 12$. V_6^* and V_{12}^* are the volume ratios in 6 and 12 dimensions, respectively. $V^*(X)$ is a function of $|X|$ in this situation. It is somewhat easier to picture things in terms of

$$R^*(X) = [V^*(X)]^{1/p},$$

which is termed by Faith (1976) the ratio of the effective radii of $C^*(X)$ and $C^o(X)$. (The effective radius of a set is the radius of a p-sphere having the same volume as the set.) In Table 2, R_6^* and R_{12}^* stand for $R^*(X)$ in 6 and 12 dimensions. In the symmetric situation $C^*(X)$ is clearly significantly smaller than $C^o(X)$.

Table 2. Volume Ratio

	$\lvert X\rvert$								
	0	1.0	2.0	4.0	6.0	8.0	10.0	20.0	50.0
V_6^*	.296	.309	.352	.561	.784	.877	.921	.980	.997
R_6^*	.816	.822	.840	.908	.960	.978	.986	.997	.999
V_{12}^*	.039	.041	.045	.075	.201	.422	.588	.881	.980
R_{12}^*	.764	.766	.772	.806	.875	.931	.957	.990	.998

Table 3 deals with the nonsymmetric situation $p = 6$, $\Sigma = I$, and A diagonal with diagonal elements $\{.1, 2, 4, 6, 8, 30\}$ (so $C = \rho^*(\Sigma+A) = (\Sigma+A)$). The entries V_i^* and R_i^* $(1 \leq i \leq 6)$ refer to the quantities $V^*(X)$ and $R^*(X)$ calculated along the ith axis. V_7^* and R_7^* are calculated along the line $|X|(1, 1, 1, 1, 1, 1)^t/6^{1/2}$. Note that

$$\mathrm{tr}(I+A\ddagger^{-1})^{-1} = 1.729 < 1.818 = 2\mathrm{ch}_{\max}(I+A\ddagger^{-1})^{-1},$$

so by Corollary 3.2.6 it must be true that $V^*(X) > 1$ for some X. Indeed for large $|X|$ along the first axis, Table 3 shows that this is the case. Such X are very unlikely to occur, however, if the prior information that θ_1 has mean 0 and variance .1 is even approximately correct.

Table 3. Volume Ratio

	$\|X\|$								
	0	1.0	3.0	5.0	7.0	9.0	11.0	20.0	50.0
V_1^*	.467	.514	.858	1.000	1.002	1.001	1.001	1.000	1.000
R_1^*	.881	.895	.975	1.000	1.000	1.000	1.000	1.000	1.000
V_2^*	.467	.476	.556	.722	.861	.919	.946	.984	.997
R_2^*	.881	.884	.907	.947	.975	.986	.991	.997	1.000
V_3^*	.467	.471	.513	.605	.732	.833	.889	.967	.995
R_3^*	.881	.882	.895	.920	.949	.970	.981	.994	.999
V_4^*	.467	.470	.498	.559	.654	.756	.833	.950	.992
R_4^*	.881	.882	.890	.908	..932	.955	.970	.991	.999
V_5^*	.467	.469	.490	.536	.608	.698	.781	.932	.989
R_5^*	.881	.882	.888	.901	.920	.942	.960	.988	.998
V_6^*	.467	.467	.473	.484	..502	.528	.560	.757	.959
R_6^*	.881	.881	.883	.886	.892	.899	.908	.955	.993
V_7^*	.467	.477	.573	.755	.901	.949	.967	.990	.998
R_7^*	.881	.884	.911	.954	.983	.991	.994	.998	1.000

For the second measure of size, $\mathrm{tr}[Q\ddagger_n(X)]^{1/2}$, general results were not obtained. However, for the case $Q = \ddagger^{-1}$ and $C = \rho\ddagger (\rho \geq 1)$ (which includes the symmetric case where Q, $\ddagger$, and C are all multiples of the identity) the following result shows that $C^n(X)$ is smaller than $C^o(X)$ if $p \geq 2$.

Theorem 3.2.7. If $Q = \Sigma^{-1}$, $C = \rho\Sigma(\rho \geq 1)$, and $p \geq 2$, then $tr[Q\hat{\Sigma}_n(X)]^{1/2} \leq tr[Q\Sigma]^{1/2}$.

Proof: Defining $a = u(||X||^2)/\rho$, it can be calculated that

$$[Q\hat{\Sigma}_n(X)]^{1/2} = [(1-u/\rho)I + wXX^t/\rho^2]^{1/2}$$

$$= (1-a)^{1/2}I + [-(1-a)^{1/2} + \{1-a+(X^t\Sigma^{-1}X)w/\rho^2\}^{1/2}](X^t\Sigma^{-1}X)\Sigma^{-1}XX^t.$$

Hence

$$tr[Q\hat{\Sigma}_n(X)]^{1/2} = (p-1)(1-a)^{1/2} + \{1-a+(X^t\Sigma^{-1}X)w/\rho^2\}^{1/2}$$

$$\leq (p-1)(1-a)^{1/2} + (1+a)^{1/2} = h(a),$$

the last step following from Lemma 3.1.1. (v). For $0 < a < 1$,

$$\frac{d}{da}h(a) = \frac{-(p-1)(1-a)^{-1/2}}{2} + \frac{(1+a)^{-1/2}}{2} < \frac{-(p-1)}{2} + \frac{1}{2} \leq 0.$$

Thus $h(a)$ is maximized at $h(0) = p = tr(Q\Sigma)^{1/2}$ and the result follows.||

Numerical calculations will not be given for the above measure of loss since (at least for the symmetric situation) $tr[Q\hat{\Sigma}_n(X)]^{1/2}$ behaves like $p(1-R^*)$.

The final measure of size is $L(X) = tr[Q\hat{\Sigma}_n(X)]$, which is also the posterior expected loss. Clearly

$$L(X) = tr(Q\Sigma) - u(||X||^2)tr(Q\Sigma C^{-1}\Sigma) + w(||X||^2)X^tC^{-1}\Sigma Q\Sigma C^{-1}X. \tag{3.23}$$

Theorem 3.2.8. $L(X) = tr[Q\hat{\Sigma}_n(X)] \leq tr[Q\Sigma]$ for all X, if and only if $tr(\Sigma Q\Sigma C^{-1}) \geq 2ch_{max}(\Sigma Q\Sigma C^{-1})$.

Proof: The "if" part follows immediately from (3.23) and the inequality

$$w(||X||^2)X^tC^{-1}\ddagger Q\ddagger C^{-1}X < \frac{2r_n(||X||^2)}{||X||^2}\left(\frac{X^tC^{-1}\ddagger Q\ddagger C^{-1}X}{X^tC^{-1}X}\right)$$

$$\leq 2u(||X||^2)ch_{max}(\ddagger Q\ddagger C^{-1}).$$

The "only if" part is proved analagously to the "only if" part of Theorem 3.2.4. Choose X to be a multiple of the eigenvector corresponding to the largest characteristic root of $C^{-1/2}\ddagger Q\ddagger C^{-1/2}$, let $|X| \to \infty$ in (3.23), and use Lemma 2.1.1 ((iii) and (ix)) and Lemma 3.1.1 (iv). ||

Corollary 3.2.9. $C^*(X)$ has smaller size (measure 3) than $C^o(X)$ for all X, if and only if

$$tr(\ddagger Q[I+A\ddagger^{-1}]^{-1}) \geq 2ch_{max}(\ddagger Q[I+A\ddagger^{-1}]^{-1}).$$

Proof: Obvious.

An interesting observation can be made concerning the relationship between $R(\theta,\delta^n)$ and $E_\theta L(X)$.

Theorem 3.2.9. If $tr(\ddagger Q\ddagger C^{-1}) \geq (2n+2)ch_{max}(\ddagger Q\ddagger C^{-1})$, then $R(\theta,\delta^n) < E_\theta L(X)$ for all θ.

Proof: Integrating by parts as in Berger (1976c) (the technique was first noticed in the symmetric case by Stein (1973)) gives

$$R(\theta,\delta^n) = tr(Q\ddagger) + E_\theta\left[\frac{-2r_n}{||X||^2}\left\{tr(\ddagger Q\ddagger C^{-1}) - \frac{2X^tC^{-1}\ddagger Q\ddagger C^{-1}X}{||X||^2}\right\}\right.$$

$$\left. - \frac{4r_n'(||X||^2)X^tC^{-1}\ddagger Q\ddagger C^{-1}X}{||X||^2} + \frac{r_n^2X^tC^{-1}\ddagger Q\ddagger C^{-1}X}{||X||^4}\right].$$

Applying Lemma 3.1.1 (iv) and (3.23) to this expression gives

$$(3.24) \qquad R(\theta,\delta^n) = E_\theta L(X) - E_\theta\Big[\frac{r_n}{||X||^2}\{tr(\sharp Q\sharp C^{-1}) - \frac{(r_n+2)X^tC^{-1}\sharp Q\sharp C^{-1}X}{||X||^2}\} + \frac{2r_n'(||X||^2)X^tC^{-1}\sharp Q\sharp C^{-1}X}{||X||^2}\Big].$$

Since $r_n'(||X||^2) > 0$, $r_n(||X||^2) < 2n$, and

$$\frac{X^tC^{-1}\sharp Q\sharp C^{-1}X}{||X||^2} \leq ch_{max}(\sharp Q\sharp C^{-1}),$$

the conclusion follows.

Corollary 3.2.10. If $\sharp Q\sharp C^{-1} = \tau I$ and $n \leq (p-2)/2$, then $R(\theta,\delta^n) < E_\theta L(X)$ for all θ.

Proof: Obvious.||

The above result was obtained for the situation $Q = \sharp = C = I$ and $n = (p-2)/2$ by Morris (1977). Stein (1974) has related results in the symmetric situation.

Theorem 3.2.9 essentially says that, under the given condition, $L(X)$ is an overestimate (on the average) of the true expected loss for δ^n. In some sense, this indicates that the corresponding confidence sets $C^n(X)$ are larger than necessary, i.e. an error on the side of conservatism is being made. Note that for the symmetric situation, $C^*(X)$ satisfies the condition of Corollary 3.2.10.

Theorem 3.2.9 is somewhat puzzling in light of the fact that if $n > p/2$ (so that the priors g_n have finite mass) then

$$\int R(\theta,\delta^n)g_n(\theta)d\theta = \int [E_\theta L(X)]g_n(\theta)d\theta.$$

(Both sides are equal to the Bayes risk, up to the normalizing constant of g_n.) If $n \leq p/2$, the integrals above are infinite, making the result of Theorem 3.2.9 possible. The following result indicates what happens for $(p-2)/2 < n \leq p/2$.

Theorem 3.2.11. If $(p-2)/2 < n \leq p/2$, then

$$\int [R(\theta,\delta^n) - E_\theta L(X)] g_n(\theta) d\theta = 0.$$

Proof: From (3.24) and Lemma 3.1.1 (iv) it can be seen that

$$\Delta(\theta) = R(\theta,\delta^n) - E_\theta L(X) = E_\theta [\frac{r_n tr(\Sigma Q \Sigma C^{-1})}{||X||^2} + \frac{t_n X^t C^{-1} \Sigma Q \Sigma C^{-1} X}{||X||^4}].$$

Denoting the integrand in the last expression above by $T(X)$, it is straightforward to check that

$$\int (E_\theta |T(X)|) g_n(\theta) d\theta < \infty$$

for $n > (p-2)/2$. Letting

$$m_n(x) = \int (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp\{-(X-\theta)^t \Sigma^{-1} (X-\theta)/2\} g_n(\theta) d\theta,$$

it is thus clear that orders of integration can be interchanged to get

$$\int \Delta(\theta) g_n(\theta) d\theta = \int T(x) m_n(x) dx.$$

Using the definitions of r_n and t_n and the fact that

$$m_n(x) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \int_0^1 \exp\{-\lambda ||x||^2/2\} \lambda^{(n-1)} d\lambda,$$

it follows that

$$\int \Delta(\theta) g_n(\theta) d\theta = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \int \{ \int \lambda^n \exp\{-\lambda ||x||^2/2\}$$
$$\times [\lambda x^t C^{-1} \Sigma Q \Sigma C^{-1} x - tr(\Sigma Q \Sigma C^{-1})] d\lambda \} dx.$$

It can be checked that the above orders of integration can be interchanged for $n > (p-2)/2$. Since

$$(2\pi)^{-p/2}\lambda\int\exp\{-\lambda x^t C^{-1}x/2\}(x^t C^{-1}\ddagger Q\ddagger C^{-1}x)dx = \lambda^{-p/2}tr(\ddagger Q\ddagger C^{-1}),$$

the resulting expression will be zero.||

The above theorem shows that for $n > (p-2)/2$, $E_\theta L(X)$ is "on the average" equal to $R(\theta,\delta^n)$, and hence $L(X)$ is not an overestimate.

In conclusion, it can be noted that for the important symmetric problem ($Q,\ddagger$, and C multiples of the identity matrix), $C^n(X)$ is smaller than $C^o(X)$ for all measures of size considered and $p \geq 2$. Even for nonsymmetric problems, $C^n(X)$ tends to be smaller than $C^o(X)$ under quite weak conditions. For example, the conditions of Theorems 3.2.4, 3.2.7, and 3.2.8 tend to be considerably weaker than the minimax condition of Theorem 2.2.1.

3.3 Probability of Coverage of C^n.

The other major facet of the confidence region C^n which is of interest is its probability of covering the true value of θ, i.e.

$$P_\theta(\theta \in C^n(X)) = \int_{\{x \in R^p:\ \theta \in C^n(x)\}} (2\pi)^{-p/2}(\det\ddagger)^{-1/2}\exp\{-(x-\theta)^t\ddagger^{-1}(x-\theta)/2\}dx. \tag{3.25}$$

Note that (3.25) is the usual (frequentist) probability of coverage, not a Bayesian probability.

Dealing with probability of coverage analytically is very difficult. It seems virtually impossible to obtain uniform (for all θ) dominance results as were obtained for size. Numerical studies of probability of coverage are very useful (and will be given), but they have the weakness in these high dimensional, many parameter settings of not being

able to adequately cover the broad range of possible problems. When discussing $R(\theta,\delta^n)$ in Section 2.2, it was shown that a very useful analytical way of determining approximate risk behavior was to look at the "tail approximation" given in lines (2.11) through (2.14). This suggests doing a similar thing for probability of coverage: obtain a large θ approximation for the probability of coverage of C^n. In looking at numerical studies, it will be seen that this approximation is a very good guide in determining the behavior of $P_\theta(\theta \in C^n(X))$.

Theorem 3.3.1. For the confidence ellipsoid

$$C^n(X) = \{\theta: \ [\theta-\delta^n(X)]^t \ddagger_n(X)^{-1}[\theta-\delta^n(X)] \leq k(\alpha)\},$$

$$P_\theta(\theta \in C^n(X)) = (1-\alpha) + \frac{2n[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}}{p\Gamma(p)\,\theta^t C^{-1}\theta}\left\{\mathrm{tr}(\ddagger C^{-1}) - \frac{(2+2n)\,\theta^t C^{-1}\ddagger C^{-1}\theta}{\theta^t C^{-1}\theta}\right\} + O(|\theta|^{-4}).$$

Proof: Given in the Appendix.||

Corollary 3.3.2. If $2n < [\mathrm{tr}(\ddagger C^{-1})/\mathrm{ch}_{\max}(\ddagger C^{-1})]-2$ and $0 < \alpha < 1$, then $P_\theta(\theta \in C_n^i(X)) > (1-\alpha)$ for large enough $|\theta|$.

Proof: Obvious from Theorem 3.3.1 and the fact that $(\theta^t C^{-1}\ddagger C^{-1}\theta)/(\theta^t C^{-1}\theta) \leq \mathrm{ch}_{\max}(\ddagger C^{-1})$.||

Corollary 3.3.3. If $C = \rho\ddagger$, then

$$P_\theta(\theta \in C^n(X)) = (1-\alpha) + \frac{2n[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}[p-(2+2n)]}{p\Gamma(p)||\theta||^2} + O(|\theta|^{-4}).$$

Proof: Obvious.||

Corollaries 3.3.2 and 3.3.3 show that $C^n(X)$ can possibly have probability of coverage greater than $(1-\alpha)$ for all θ only if $n \leq (p-2)/2$. Unfortunately,

the estimator δ^n is inadmissible if $n < (p-2)/2$. Thus to obtain a good estimator and a probability of coverage which is not seriously worse than $(1-\alpha)$, it seems that the choice of $n = (p-2)/2$ should be made. In part, this is why δ^* and C^* were recommended with the choice $n = (p-2)/2$.

For problems in which $C \neq \rho\sharp$, Thoerem 3.3.1 is useful in determining the directions in which $C^*(X)$ has greater or smaller probability of coverage than $(1-\alpha)$. Indeed, being as the error term in Theorem 3.3.1 is $0(|\theta|^{-4})$ while the dominant terms is $0(|\theta|^{-2})$, the approximation is fairly accurate for even moderate values of $|\theta|$. (Numerical studies showed this to be the case.)

As an example, the case $p = 6$, $\sharp = I$, A diagonal with diagonal elements $\{.1, 2, 4, 6, 8, 30\}$, and $1 - \alpha = .90$ was considered. (This example was discussed in Section 3.2 with respect to the size of $C^*(X)$.) The probabilities of coverage $P_\vartheta(\theta \in C^*(X))$ were calculated along the six axes and along the first quadrant diagonal. Table 4 gives the results for various values of $|\theta|$. (p_i stands for the probability of coverage along the ith axis ($1 \leq i \leq 6$), while p_d is for along the diagonal.) From Theorem 3.3.1 (with $C = \rho^*(\sharp+A)$) it could be predicted that C^* would have a probability of coverage smaller than $(1-\alpha)$ for large enough $|\theta|$ along the first two axes and the diagonal, and probability of coverage larger than $(1-\alpha)$ along the remaining axes. This behavior is exactly what is observed in Table 4. The coverage probability along the first axis appears particularly bad, until it is remembered that the prior input states that θ_1 has mean zero and variance .1. Hence for θ_1 within several standard deviations of zero, the probability of coverage is greater than $(1-\alpha)$.

Table 4. Probabilities of Coverage of C*(X).

	$\|\theta\|$									
	0	1.0	1.5	2.0	3.0	4.0	5.0	6.0	10.0	15.0
p_1	.960	.935	.885	.820	.787	.821	.852	.870	.890	.895
p_2	.960	.957	.953	.947	.931	.915	.903	.897	.899	.899
p_3	.960	.959	.956	.955	.950	.943	.933	.923	.908	.904
p_4	.960	.959	.958	.957	.952	.948	.941	.935	.914	.905
p_5	.960	.960	.960	.959	.956	.953	.949	.944	.923	.910
p_6	.960	.960	.960	.960	.960	.959	.959	.958	.953	.943
p_d	.960	.956	.950	.941	.911	.873	.850	.848	.880	.892

For symmetric problems (or more generally those with $C = \rho\ddagger$), one would hope that $C^*(X)$ does have coverage probability greater than $(1-\alpha)$. Unfortunately, Theorem 3.3.1 or Corollary 3.3.3. are no longer of any assistance, since $[p-(2+2n)] = 0$. It is thus the term of order $|\theta|^{-4}$ that is dominant, as the following theorem shows.

Theorem 3.3.4. If $C = \rho\ddagger$ and $n = (p-2)/2$, then

$$P_\theta(\theta \in C^n(X)) = (1-\alpha) + \frac{(p-2)[k(\alpha)/2]^{p/2}\exp\{-k(\alpha)/2\}}{4p\Gamma(p/2)(\theta^t\ddagger^{-1}\theta)^2}\{4p(p-2)$$

$$+\frac{k(\alpha)}{2(p+2)}[p^3+2p^2-32p-48]\} + O(|\theta|^{-6}).$$

Proof: Given in the Appendix.||

Corollary 3.3.5. If $A = \rho\ddagger$, then for large $|\theta|$, $P_\theta(\theta \in C^*(X)) > (1-\alpha)$ providing

(i) $0 < k(\alpha) \leq 1.212$ (i.e. $0 < (1-\alpha) < .25$) when $p = 3$,

(ii) $0 < k(\alpha) < 4.8$ (i.e. $0 < (1-\alpha) < .69$) when $p = 4$,

(iii) $0 < k(\alpha) < 25.45$ (i.e. $0 < (1-\alpha) < .9999$) when $p = 5$,

(iv) $0 < k(\alpha) < \infty$ (i.e. $0 < (1-\alpha) < 1$) when $p \geq 6$.

Proof: The conditions on $k(\alpha)$ are simply those for which $\{4p(p-2) + k(\alpha)[2(p+2)]^{-1}[p^3+2p^2-32p-48]\} > 0$. Theorem 3.3.4 thus gives the desired result.||

For $p \geq 6$ (and virtually always for $p = 5$), the coverage probability of C^* is thus greater than $(1-\alpha)$ for large enough $|\theta|$. To determine the behavior for small $|\theta|$, numerical studies were conducted for $(1-\alpha) = .90$, $p = 4$, 6, and 12, and $C = 2\ddagger$(which would result, say, if $A = \ddagger$). The results are given in Table 5 for various values of $(\theta^t\ddagger^{-1}\theta)^{1/2}$. The coverage probability is never much worse than .90, and for small $(\theta^t\ddagger^{-1}\theta)^{1/2}$ was considerably better. Note that as predicted by Corollary 3.3.5, the coverage probability fell below .90 for $p = 4$ and large $|\theta|$, but was above .90 for $p = 6$ and 12 and large $|\theta|$. The dip below .90 at $(\theta^t\ddagger^{-1}\theta)^{1/2} = 8$ and $p = 12$ is somewhat surprising. The $O(|\theta|^{-4})$ term of Theorem 3.3.4 is apparently not yet dominant at this point. (The probabilities were computed by simulation, using 60,000 generations of X.)

Table 5. Probabilities of Coverage of $C^*(X)$

	$(\theta^t\ddagger^{-1}\theta)^{1/2}$									
	0	1	2	3	4	5	6	8	10	15
4	.971	.965	.945	.918	.902	.897	.898	.898	.898	.898
p 6	.993	.989	.976	.946	.916	.902	.900	.901	.901	.901
12	1.000	.999	.998	.988	.958	.921	.900	.895	.898	.900

3.4. Comparison With Other Confidence Procedures

As mentioned at the beginning of Section 3, several other multivariate confidence procedures have been proposed. For the most part they have been presented and studied only in the symmetric situation (Q, Σ, and A multiples of I), so the comparisons in this section will be restricted to that case. Along with $C^{O}(X)$ and $C^{*}(X)$, we will consider

$$C^{B\text{-}J}(X) = \{\theta: |\theta-\delta^{*}(X)|^{2} \leq k(\alpha)\},$$

and

$$C^{M}(X) = \{\theta: [\theta-\delta^{*}(X)]^{t}\Sigma_{M}^{-1}[\theta-\delta^{*}(X)] \leq k(\alpha)\},$$

where $\Sigma_{M}(X)$ consists of the diagonal elements of $\Sigma^{*}(X)$. $C^{B\text{-}J}(X)$ is simply the usual confidence region centered at the improved estimator δ^{*} (in the spirit of the Brown (1966) and Joshi (1967) confidence sets). $C^{M}(X)$ is related to the region suggested by Morris (1977) in the symmetric situation. One difference is that his choice of C in the prior g_{n} was C = I, not C = ρ^{*}I as proposed here. (Some comments about both choices will be made.) The major difference is that confidence intervals, not confidence ellipsoids, are considered in Morris (1977). Hence overall probability of coverage is not the goal he pursues. To make meaningful comparisons, therefore, an ellipsoid using the variances in Morris (1977) is considered.

The other major proposed confidence regions, those of Stein (1963) and (1974) and Faith (1976), will not be discussed. Stein's regions are developed heuristically for large p and without modification are probably not suitable for small p. Faith's regions will not be considered for two reasons. First, as they are developed in a Bayesian fashion (though in the symmetric case), their performance is quite likely very similar to $C^{*}(X)$. On the other hand, they have a complicated shape and are hard to work with or evaluate. The relative simplicity of the other procedures makes them attractive.

In comparing sizes, only volume will be discussed, though similar conclusions hold for other measures of size. Since $C^{O}(X)$ and $C^{B-J}(X)$ have the same size, the results of Section 3.2 hold for both. (See in particular Corollary 3.2.5 and Table 2.) $C^{*}(X)$ clearly achieves a very significant reduction in size over $C^{O}(X)$ or $C^{B-J}(X)$. The following theorem also shows that $C^{*}(X)$ has smaller volume than $C^{M}(X)$.

Theorem 3.4.1. Assume C and Σ are diagonal, and let $\Sigma_n^M(X)$ denote the matrix of diagonal elements of $\Sigma_n(X)$. Then $\det[\Sigma_n(X)]^{1/2} \leq \det[\Sigma_n^M(X)]^{1/2}$.

Proof: Define

$$z_i = w(||X||^2)x_i^2\sigma_i^2/[c_i^2\{1-u(||X||^2)\sigma_i^2/c_i\}],$$

where $\{\sigma_i^2\}$ and $\{c_i\}$ are the diagonal elements of Σ and C. A calculation gives that

$$\det\Sigma_n^M(X) = [\det\Sigma]\,[\det\{I-u(||X||^2)C^{-1}\Sigma\}]\,[\prod_{i=1}^{p}(1+z_i)]. \tag{3.26}$$

Noting from Lemma 3.2.2 that

$$\det\Sigma_n(X) = [\det\Sigma][\det\{I-u(||X||^2)C^{-1}\Sigma\}][1+\sum_{i=1}^{p} z_i], \tag{3.27}$$

the conclusion follows from the inequality

$$(1 + \sum_{i=1}^{p} z_i) \leq \prod_{i=1}^{p}(1+z_i)\cdot||$$

As an aside, it is interesting to note that (3.26) and Lemma 3.2.3 can be used to show that $\det[\Sigma_n^M(X)] \leq \det[\Sigma]$ if $\mathrm{tr}(C^{-1}\Sigma) \geq 2\mathrm{ch}_{\max}(C^{-1}\Sigma)$. (The

proof is essentially given in Theorem 3.2.4 starting with (3.22).) Hence for the symmetric situation (and $p \geq 2$), $C^M(X)$ has smaller volume than $C^O(X)$ or $C^{B\text{-}J}(X)$.

Though Theorem 3.4.1 shows that $C^*(X)$ has smaller volume than $C^M(X)$, the difference is much less than that between $C^*(X)$ and $C^O(X)$. From (3.26) and (3.27) it is indeed clear that if X lies along one of the axes, then $\det[\ddagger_n^M(X)] = \det[\ddagger_n(X)]$. When X lies on a diagonal, on the other hand, the difference is greatest (for the symmetric situation). Table 6 gives the values of $[\det\ddagger^*(X)/\det\ddagger^M(X)]^{1/2}$ along the diagonals for various values of X, p, and C when $\ddagger = I$. Morris always chooses C = I, while C = 2I is more typical of $C = \rho^*(\ddagger+A)$ suggested here.

Table 6. Volume Ratio of $C^*(X)$ and $C^M(X)$

	$\lvert X\rvert$								
p	1	2	3	4	5	6	7	9	15
12(C=2I)	1.000	.999	.995	.981	.949	.921	.929	.970	.996
6(C=2I)	1.000	.997	.986	.971	.970	.980	.989	.996	.999
6(C=I)	.990	.912	.848	.900	.955	.978	.989	.996	.999

To compare probabilities of coverage, numerical studies were conducted. Tables 7, 8, and 9 give results for $\ddagger = I$, C = 2I, and p equal 4, 6, and 12 respectively. Both C^* and $C^{B\text{-}J}$ have probabilities of coverage which depend only on $|\theta|$. C^M, on the other hand, does not. Hence results for C^M are given for θ along the axes (C_a^M) and for θ along the diagonals (C_d^M).

Table 7. Probabilities of Coverage (p=4).

	$\|\theta\|$									
	0	1	2	3	4	5	6	8	10	15
C^*	.971	.965	.945	.918	.902	.897	.897	.898	.898	.899
C^{B-J}	.970	.967	.959	.945	.928	.914	.908	.903	.902	.900
C_a^M	.959	.954	.940	.921	.909	.904	.902	.900	.899	.899
C_d^M	.959	.954	.938	.916	.900	.895	.895	.898	.898	.899

Table 8. Probabilities of Coverage (p=6).

	$\|\theta\|$									
	0	1	2	3	4	5	6	8	10	15
C^*	.993	.989	.976	.946	.916	.902	.900	.901	.901	.901
C^{B-J}	.990	.989	.985	.976	.962	.944	.930	.917	.912	.906
C_a^M	.981	.977	.965	.945	.927	.917	.912	.907	.904	.902
C_d^M	.981	.977	.964	.939	.912	.898	.895	.899	.901	.901

Table 9. Probabilities of Coverage (p=12).

	$\|\theta\|$									
	0	1	2	3	4	5	6	8	10	15
C^*	1.000	.999	.998	.988	.958	.921	.900	.895	.898	.900
C^{B-J}	.999	.999	.999	.997	.995	.990	.978	.952	.936	.917
C_a^M	.995	.994	.991	.980	.961	.942	.933	.922	.912	.903
C_d^M	.995	.994	.991	.979	.951	.916	.893	.888	.894	.898

Except for small $|\theta|$, C^{B-J} has better probability of coverage than C*. On the other hand, C* has significantly smaller volume than C^{B-J} (Table 2). In looking at the tradeoffs involved, the smaller size seems to more than offset the smaller probability of coverage. From an applications viewpoint, the confidence procedure, C*, seems more appropriate also. It can be reported as a $(1-\alpha)$ confidence region and will have a definitely reportable smaller size than $C^{O}(X)$. $C^{B-J}(X)$, on the other hand, has the same size as $C^{O}(X)$ and can also only be reported as a $(1-\alpha)$ confidence region. The gains in probability of coverage if the true θ happens to be small are hard to report. C^{B-J} would, in a conservative sense, be more competitive in nonsymmetric situations, since its probability of coverage would be less likely to drop below $(1-\alpha)$ than would the probability of coverage of C*.

C* and C^{M} have very similar probabilities of coverage. (Note that both are calculated at $C = 2I$ for comparison purposes. The choice of $C = I$ gives less attractive results for both regions.) C^{M} is better along the axes, while C* is better along the diagonals. The smaller size of C*(X) and its greater simplicity in nonsymmetric situations make it attractive. Both procedures, however, should do quite well.

Section 4. Incorporation of Prior Information.

As mentioned in Section 1, prior input in the form of a prior mean vector μ and a prior covariance matrix A is envisaged. The use of A in the development of δ^* and C* has already been discussed. To use μ, the estimator and confidence region should be centered at μ. Thus

$$\delta^*(X) = X - \frac{r^*(||X-\mu||^2)\Sigma(\Sigma+A)^{-1}(X-\mu)}{(X-\mu)^t(\Sigma+A)^{-1}(X-\mu)} \tag{4.1}$$

is the recommended estimator. The definition of $\xi^*(X)$ is unchanged except that X should be replaced by $X-\mu$ in all expressions. This shift changes none of the properties or results established in Sections 2 and 3.

It is sometimes desirable to choose C^{-1} to be singular. The only change which should then be made in the definitions of δ^n and C^n is to choose $n = ([\text{rank } C^{-1}]-2)/2$ instead of $n = (p-2)/2$. The rank of C^{-1} is the effective dimensionality of the problem. This can be seen by diagonalizing ξ and C, and then noting that δ^n and ξ^n are the generalized Bayes estimator and posterior covariance matrix for a subproblem of rank C^{-1}. Thus all the results of Section 2 and 3 (with the exception of the admissibility of δ^n) hold with p replaced by $[\text{rank } C^{-1}]$.

The reason for choosing C^{-1} singular would be that in some directions you have no prior information whatsoever (or alternatively, that A has infinite characteristic roots in those directions). The corresponding coordinates are then effectively excluded from the correction term of the estimator δ^n, and are dealt with by the usual estimator δ^o and confidence region C^o.

An example of the use of singular C^{-1} is when shrinkage towards the common mean $\bar{\bar{X}} = \sum_{i=1}^{p} X_i/p$ is desired. Defining (1) as the matrix of all ones, $\bar{1}$ as the column vector of ones, and letting $C^{-1} = I - \frac{1}{p}(1)$, it is easy to check that (for $\xi = I$)

$$\delta^*(X) = X - \frac{r_n(|X-\bar{X}\bar{1}|^2)(X-\bar{X}\bar{1})}{|X-\bar{X}\bar{1}|^2}, \tag{4.2}$$

an estimator which shrinks towards the common mean. Note that C^{-1} has rank (p-1), so $n = (p-3)/2$ is the choice of n in r_n above. Choosing C^{-1} as above is essentially a statement that the θ_i are felt to be similar (or their priors have a common mean or their prior is exchangeable), but that the common value that

the θ_i are near is totally unknown. This last assumption seems somewhat extreme intuitively, and the following Bayesian considerations suggest a reasonable alternative.

Assume that the θ_i are thought to be a random sample from a normal distribution with mean θ_0 and variance σ^2. (It is convenient to develop u and A through the assumption of normal priors due to their ease of manipulation.) It is often assumed that θ_0 also has a normal distribution with mean μ_0 and variance σ_0^2. (This problem is discussed in Lindley (1972), where earlier works on the model are also referenced.) As pointed out in Lindley (1972), this two stage prior is equivalent to assuming that θ has a p-variate normal distribution with mean $\mu = \mu_0\bar{1}$ and covariance matrix $A = (\sigma^2 I + \sigma_0^2(1))$. The common Bayesian technique is to use the linear Bayes estimator, letting $\sigma_0^2 \to \infty$. (The prior information at the second stage is deemed vague, so taking σ_0^2 to infinity results in a more robust estimator.) Due to the fact that δ^* is already quite robust, however, the best guesses u and A can safely be used directly in δ^*. There is no need to let $\sigma_0^2 \to \infty$. Note that at the two extremes, letting $\sigma_0^2 \to \infty$ in δ^* would result in (4.2) (providing $\Sigma = I$ and $n = (p-3)/2$ were used), while choosing $\sigma_0^2 = 0$ would simply result in an estimator shrinking towards the believed mean $\mu_0\bar{1}$..

As another example of the use of prior information, assume that the linear restriction

$$B(\theta - \theta_0) = 0$$

is thought to hold, where θ_0 is a p vector and B is a (k×p) matrix ($k \leq p$) of rank k. Suppose a (k×k) positive definite matrix A is also determined, where A reflects the accuracy with which the linear restrictions are believed

to hold. (A can be thought of as the estimated covariance matrix of the prior distribution of $B(\theta - \theta_0)$.)

The simplest way to proceed is to define $Y = B(X - \theta_0)$, $\eta = B(\theta - \theta_0)$, and $S = Q^{-1}B^t(BQ^{-1}B^t)^{-1}$. (Q is from the loss function.) Theorem 2 of Berger and Bock (1977) states that the improvement (over δ^o) in estimating η by $\delta(X) = X - S\gamma(B(X - \theta_0))$ is equal to the improvement (over δ^o) in estimating η by $\bar{\delta}(Y) = Y - \gamma(Y)$ under the quadratic loss $(\delta - \eta)^t(BQ^{-1}B^t)^{-1}(\delta - \eta)$. (The problem decomposes into the estimation of $[I-SB]\theta$ by $[I-SB]X$, and η by $\bar{\delta}(Y)$.)

The obvious choice for $\gamma(Y)$ is

$$\gamma(Y) = \frac{r_n(||Y||^2)\, \Sigma_Y C^{-1} Y}{||Y||^2},$$

where $\Sigma_Y = B\Sigma B^t$, $C = \rho^*(\Sigma_Y + A)$, and $n = (k-2)/2$. A reasonable choice for the confidence region is simply the usual confidence ellipsoid for $(I-SB)\theta$, and $C^*(Y)$ for η. It is at first sight disturbing to estimate $(I-SB)\theta$ by the usual estimator $(I-SB)X$, when the dimension (i.e. rank $([I-SB]) = p-k$) could be three or more. In choosing a "k-dimensional" linear restriction, however, it is essentially being said that there is no prior information about the remaining dimensions (i.e. about $[I-SB]\theta$). If indeed this is the case, then it is a waste of time trying to adjust the usual estimator $(I-SB)X$ of $(I-SB)\theta$, since the "chance" of θ being such that there is significant improvement is negligible.

Section 5. Unknown Variance.

In applications, it is important to consider the situation where the covariance matrix of X is unknown. Attention will be restricted to the case

where the covariance matrix is of the form $\sigma^2\Sigma$, Σ known but σ^2 unknown. (This is the common situation in regression problems.)

There are two possible appraoches to dealing with this problem. The first is simply to replace σ^2 by an estimate in δ^* and C^* (with appropriate changes to $k(\alpha)$ in C^*). The second is to place a prior upon σ^2 (in addition to θ), and try to develop δ^* and C^* in terms of the combined prior information.

The second approach was used by Strawderman (1973) for the case $\Sigma = I$ (in g_n). (M. E. Bock (personal communication) has been able to explicitely evaluate the resulting estimator.) Unfortunately, the resulting estimator is extremely complex, even in this simple setting. The problems of constructing such an estimator for the nonsymmetric setting, and then of meaningfully analyzing it, seem considerable. Indeed the priors placed on σ^2 are rather unintuitive, and whether or not they have a beneficial effect on the estimaor is unclear. It should be emphasized that δ^* and C^* were developed in a Bayesian fashion mainly because it appeared necessary to use prior information in the choice of a competitor to $\delta^0(X) = X$. There is no such compelling reason to use prior information on σ^2 in constructing δ^*. The approach that will be adopted is thus the first approach, merely replacing σ^2 by an estimate in σ^* and C^*. (Of course, if significant prior information about σ^2 were available, it would be reasonable to use this in the estimation of σ^2, but this could be left up to individual taste. Note that the effect upon δ^* would probably be slight, in the sense that δ^* would still probably be very robust, but the effect of wrong prior information about σ^2 on C^* could be considerable.)

When σ^2 is unknown, assume a random variable S^2 is observable, where S^2/σ^2 has a chi square distribution with m degrees of freedom. A suitable estimate of σ^2 for use in δ^* and C^* is $S^2/(m+2)$. Thus $[S^2/(m+2)]\Sigma$ and

$C = \rho^*\{[S^2/(m+2)]\ddagger + A\}$ should be used in δ^* and C^* in place of the previous $\ddagger$ and C. A reason for choosing $S^2/(m+2)$ as the estimator of σ^2 is that it is the natural estimator for certain minimax type results. The following theorem is an example. For convenience, define

$$(5.1)\quad G(Q,\ddagger,A) = \lim_{t\to 0} \frac{tr[(\ddagger Q\ddagger(t\ddagger+A)^{-1}]}{ch_{max}[\ddagger Q\ddagger(t\ddagger+A)^{-1}]}.$$

Note that if A is nonsingular, then $G(Q,\ddagger, A) = tr(\ddagger Q\ddagger A^{-1})/ch_{max}(\ddagger Q\ddagger A^{-1})$.

Theorem 5.1. Assume Q^{-1}, $\ddagger$, and A are simultaneously diagonalizable, with resulting diagonal elements $\{q_i^{-1}\}$, $\{d_i\}$, and $\{A_i\}$, satisfying

$$(5.2)\quad [(A_j/d_j) - (A_i/d_i)](d_iq_i - d_jq_j) \geq 0, \quad 1 \leq i, j \leq p.$$

Let $C = \rho(S^2)(\frac{S^2}{(m+2)}\ddagger + A)$, where ρ is nondecreasing in S^2. Then

$$\delta^n(X,S^2) = (I - \frac{r_n(||X||^2)S^2\ddagger C^{-1}}{|||X||^2(m+2)})X$$

has smaller risk than $\delta^o(X) = X$, providing $n \leq G(Q,\ddagger,A) - 2$.

Proof: Integrating by parts as in Theorem 3.2.9 gives

$$(5.2)\quad R(\theta,\sigma^2,\delta^n)-R(\theta,\sigma^2,\delta^o)=E_{\theta,\sigma^2}[\frac{-2r_nS^2\sigma^2}{||X||^2(m+2)}\{tr(\ddagger Q\ddagger C^{-1})-\frac{2(X^tC^{-1}\ddagger Q\ddagger C^{-1}X)}{||X||^2}\}$$

$$-\frac{4\sigma^2r_n'(||X||^2)S^2(X^tC^{-1}\ddagger Q\ddagger C^{-1}X)}{||X||^2(m+2)}+E_{\theta,\sigma^2}[\frac{r_n^2(X^tC^{-1}\ddagger Q\ddagger C^{-1}X)S^4}{||X||^4(m+2)^2}].$$

Efron and Morris (1976) proved the identity

$$(5.4)\quad E_{\sigma^2}[g(S^2)S^2] = \sigma^2m\,E_{\sigma^2}[g(S^2)] + 2\sigma^2\,E_{\sigma^2}[S^2g'(S^2)],$$

for any differentiable function g for which the expectations exist. Defining

$$h(S^2) = (X^tC^{-1}\sharp Q\sharp C^{-1}X)/||X||^4,$$

(recall C is a function of S^2) and setting

$$g(S^2) = r_n^2(||X||^2)S^2h(S^2)/(m+2)^2,$$

it follows from (5.4) that

$$(5.5)\qquad E[\frac{r_n^2S^4h(S^2)}{(m+2)^2}] = \sigma^2 m\, E[\frac{r_n^2(S^2h(S^2)}{(m+2)^2}]$$

$$+ 2\sigma^2E[S^2\{\frac{2r_n(\frac{d}{dS^2}r_n(||X||^2))S^2h(S^2)}{(m+2)^2} + \frac{r_n^2h(S^2)}{(m+2)^2} + \frac{r_n^2S^2h'(S^2)}{(m+2)^2}\}].$$

From the definition of C and the assumption that ρ is nondecreasing in S^2, it is clear that $||X||^2 = X^tC^{-1}X$ is nonincreasing in S^2. Hence

$$(5.6)\qquad \frac{d}{dS^2}r_n(||X||^2) \leq 0.$$

Define $Y = TX$, where T is a $(p\times p)$ matrix such that $TQ^{-1}T^t$, $T\sharp T^t$, and TAT^t are all diagonal matrices. It is easy to check that

$$h(S^2) = [\sum_{i=1}^{p} Y_i^2q_id_i/(\frac{S^2}{m+2} + \frac{A_i}{d_i})^2]/[\sum_{i=1}^{p} Y_i^2/(\frac{S^2}{m+2} + \frac{A_i}{d_i})]^2.$$

Defining $b_i = [S^2/(m+2)] + [A_i/d_i]$, a calculation gives that

$$h'(S^2) = \frac{-2}{(m+2)\left(\sum_{i=1}^{p} Y_i^2/b_i\right)^3}\left\{\left(\sum_{i=1}^{p} Y_i^2 q_i d_i/b_i^3\right)\left(\sum_{j=1}^{p} Y_j^2/b_j\right) - \left(\sum_{i=1}^{p} Y_i^2 q_i d_i/b_i^2\right)\left(\sum_{j=1}^{p} Y_j^2/b_j^2\right)\right\}$$

$$= \frac{-2}{(m+2)\left(\sum_{i=1}^{p} Y_i^2/b_i\right)^3}\left\{\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[b_j^2 q_i d_i - b_i b_j q_i d_i\right]\right\}$$

$$= \frac{-2}{(m+2)\left(\sum_{i=1}^{p} Y_i^2/b_i\right)^3}\left\{\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[b_j^2 q_i d_i - b_i b_j q_i d_i + b_i^2 q_j d_j - b_i b_j q_j d_j\right]\right\}$$

$$= \frac{-2}{(m+2)\left(\sum_{i=1}^{p} Y_i^2/b_i\right)^3}\left\{\sum_{i=1}^{p}\sum_{j=1}^{(i-1)} \frac{Y_i^2 Y_j^2}{b_i^3 b_j^3}\left[(b_j - b_i)(b_j q_i d_i - b_i q_j d_j)\right]\right\}.$$

Using the definition of b_i, a calculation gives that

$$(5.7)\qquad (b_j - b_i)(b_j q_i d_i - b_i q_j d_j) = \left(\frac{A_j}{d_j} - \frac{A_i}{d_i}\right)(q_i d_i - q_j d_j)\frac{S^2}{(m+2)} + \left(\frac{A_j}{d_j} - \frac{A_i}{d_i}\right)\left(\frac{A_j q_i d_i}{d_j} - \frac{A_i q_j d_j}{d_i}\right).$$

The first term on the right hand side of (5.7) is nonnegative by (5.2). The second term is nonnegative since (5.2) implies that the two factors of the second term have the same sign (or one is zero). It can thus be concluded that $h'(S^2) \leq 0$. Together with (5.5) and (5.6), this implies that

$$(5.8)\qquad E_{\sigma^2}\left[\frac{r_n^2 S^4 h(S^2)}{(m+2)^2}\right] \leq \sigma^2 E_{\sigma^2}\left[\frac{r_n^2 S^2 h(S^2)}{(m+2)}\right].$$

Using (5.3), (5.8) and the facts that $r_n'(||X||^2) > 0$, $r_n(||X||^2) < 2n$, and $(X^tC^{-1}\ddagger Q\ddagger C^{-1}X)/||X||^2 \leq ch_{max}(\ddagger Q\ddagger C^{-1})$, it follows that

$$(5.9)\quad R(\theta,\sigma^2,\delta^n) - R(\theta,\sigma^2,\delta^o) < E_{\theta,\sigma^2}\Big[\frac{-2r_n S^2\sigma^2 ch_{max}(\ddagger Q\ddagger C^{-1})}{||X||^{2(m+2)}} \times \Big\{ \frac{tr(\ddagger Q\ddagger C^{-1})}{ch_{max}(\ddagger Q\ddagger C^{-1})} - (2+n)\Big\}\Big].$$

Clearly

$$(5.10)\quad \frac{tr(\ddagger Q\ddagger C^{-1})}{ch_{max}(\ddagger Q\ddagger C^{-1})} = \sum_{i=1}^{p} \frac{(d_i q_i/b_i)}{\max_j\{d_j q_j/b_j\}} \leq 1 + \sum_{i\neq k} \frac{b_k d_i q_i}{b_i d_k q_k},$$

where k is the coordinate at which the maximum is attained. But if $d_k q_k/b_k \geq d_i q_i/b_i$ for $i \neq k$, then for (5.2) to hold it must be true that $b_k \leq b_i$, or equivalently that $A_k/d_k \leq A_i/d_i$. Hence

$$\frac{b_k}{b_i} = \frac{S^2/(m+2) + A_k/d_k}{S^2/(m+2) + A_i/d_i}$$

in nondecreasing in S^2. It follows that (5.10) is minimized at $S^2 = 0$, attaining the value $G(Q,\ddagger,A)$. Together with (5.9), this establishes that

$$R(\theta,\sigma^2,\delta^n) - R(\theta,\sigma^2,\delta^o) < -E_\theta\Big[\frac{2r_n S^2\sigma^2 ch_{max}(\ddagger Q\ddagger C^{-1})}{||X||^{2(m+2)}} \{G(Q,\ddagger,A) - (2+n)\}\Big].$$

By the condition on n, the argument of the expectation is positive and the conclusion follows.||

Two special cases of interest are given in the following corollaries.

Corollary 5.2. If $Q = \tau\Sigma^{-1}$, then δ^n (chosen as in Theorem 5.1) has smaller risk than δ^o if $n \leq G(Q,\Sigma,A) - 2$. (If A is nonsingular, $G(Q,\Sigma,A) = tr(\Sigma A^{-1})/ch_{max}(\Sigma A^{-1})$.)

Proof: Clearly $Q^{-1} = \tau^{-1}\Sigma$, Σ, and A are simultaneously diagonalizable. Also, $d_i q_i = \tau$ for all i, so that (5.2) is satisfied. The conclusion follows from Theorem 5.1.||

Note that $Q = \Sigma^{-1}$ is an often considered choice of Q, as it gives rise to a loss which is invariant and more importantly is the natural loss for the prediction problem of linear regression. (Predict the value of a future observation arising from the same design matrix.)

Corollary 5.3. If $A = \tau\Sigma$, then δ^n (chosen as in Theorem 5.1) has smaller risk than δ^o if $n \leq [tr(\Sigma Q)/ch_{max}(\Sigma Q)] - 2$.

Proof: Q^{-1}, Σ, and $A = \tau\Sigma$ are all simultaneously diagonalizable and $A_i/d_i = \tau$ for all i. Hence (5.2) is satisfied and Theorem 5.1 can be applied to give the desired result.||

The estimator δ^n is undoubtedly uniformly better than δ^o in situations where (5.2) is not satisfied, but a more general proof was not found. Note, in any case, from the statement of Theorem 5.1, that m (the degrees of freedom of S^2) is not part of the condition of the theorem. This is why $S^2/(m+2)$ seemed the natural estimator of σ^2 to use in δ^*.

Corollary 5.4. If Q^{-1}, Σ and A are simultaneously diagonalizable and satisfy (5.1), then δ^* has smaller risk than δ^o if $p \leq 2\,G(Q,\Sigma,A) - 2$.

Proof: Obvious from Theorem 5.1 since ρ^* as chosen in (2.10) is nondecreasing in S^2.||

The estimation of σ^2 does not affect the robust Bayesian properties of δ^* appreciably, so numerical studies (such as Table 1) will not be presented for this case.

When estimating σ^2 by $S^2/(m+2)$, the appropriate definition of the confidence region C^n is now

$$C^n(X,S^2) = \{\theta: \ [\theta-\delta^n(X,S^2)]^t \Sigma_n(X,S^2)^{-1}[\theta-\delta^n(X,S^2)] \leq k(\alpha)\},$$

where δ^n and Σ_n are defined as earlier with Σ replaced by $[S^2/(m+2)]\Sigma$ and $k(\alpha) = (m+2)p\, F_{p,m}(1-\alpha)/m$, $F_{p,m}(1-\alpha)$ being the $100(1-\alpha)$th percentile of the F distribution with p and m degress of freedom.

In considering the size of $C^n(X,S^2)$, the results in Section 3.2 all hold with Σ replaced by $[S^2/(m+2)]\Sigma$. The conditions of the theorems then depend on S^2, however, at least for $C^*(X,S^2)$ which chooses $C = \rho^*([S^2/(m+2)]\Sigma + A)$. Global theorems can be developed, if desired, an example of which is the following. Note that the usual confidence ellipsoid when σ^2 is unknown is

$$C^o(X,S^2) = \{\theta: \ (\theta-X)^t \Sigma^{-1}(\theta-X) \leq (\frac{S^2}{m})p\, F_{p,m}(1-\alpha)\}.$$

Theorem 5.5. $C^*(X,S^2)$ has smaller volume than $C^o(X,S^2)$ for all X and S^2 if $G(\Sigma^{-1},\Sigma,A) \geq 2$. (G is defined in (5.1).)

Proof: By Corollary 3.2.6, it is only necessary to show that for all $S^2 > 0$,

$$(5.11) \qquad \frac{\mathrm{tr}[I+A\Sigma^{-1}/\{S^2/(m+2)\}]^{-1}}{\mathrm{ch}_{\max}[I+A\Sigma^{-1}/\{S^2/(m+2)\}]^{-1}} \geq 2.$$

Letting $\{b_i\}$ denote the roots of $\Sigma^{-1/2}A\Sigma^{-1/2}$, it is clear that (5.11) can be rewritten

$$(5.12) \qquad \sum_{i=1}^{p} \frac{S^2/(m+2) + \min\{b_j\}}{S^2/(m+2) + b_i} \geq 2.$$

The expression on the left hand side of (5.12) is clearly minimized as $S^2 \to 0$. But the limit as $S^2 \to 0$ is nothing but $G(\Sigma^{-1},\Sigma,A)$, and the conclusion follows.||

Tables 2 and 3 still give typical volume ratios of $C^*(X,S^2)$ to $C^o(X,S^2)$ (when $S^2 = m + 2$ for example).

No attempt was made to determine the large θ approximation to $P_\theta(\theta \in C^*(X,S^2))$, since the results are likely to be similar to those of Section 3.3. Numerical studies were performed, however, the results being given in Tables 10, 11, and 12. Table 10 gives $P_\theta(\theta \in C^*(X,S^2))$ for $p = 6$, $\Sigma = A = I$, $\sigma^2 = 1$, and $m = 10$ and 15. (p_{10} and p_{15} are the values when $m = 10$ and $m = 15$, respectively.) Tables 11 and 12 give $P_\theta(\theta \in G^*(X,S^2))$ for $p = 6$, $\Sigma = A = I$, $m = 15$, and σ^2 equal to .2 and 5 respectively. The results in the tables are all quite satisfactory.

In conclusion, it appears that the estimation of σ^2 does not seriously reduce the benefits of using δ^* and C^*.

Table 10. Probabilities of Coverage, $\sigma^2 = 1$.

	$\lvert\theta\rvert$									
	0	1	2	3	4	5	6	8	10	15
p_{15}	.973	.968	.949	.924	.907	.901	.900	.900	.900	.900
p_{10}	.962	.956	.940	.923	.910	.905	.903	.902	.901	.900

Table 11. Probabilities of Coverage, $\sigma^2 = .2$.

	$\lvert\theta\rvert$								
	0.0	.40	1.0	1.5	2.0	3.0	5.0	10.0	15.0
p	.941	.941	.938	.933	.926	.914	.906	.902	.901

Table 12. Probabilities of Coverage, $\sigma^2 = 5$.

	$\lvert\theta\rvert$									
	0	2	4	6	10	15	20	25	30	50
p	.978	.974	.960	.930	.904	.904	.903	.902	.901	.901

Section 5. Generalizations and Comments

1. An interesting feature of δ^n can be observed using Lemma 2.1.1 (v), namely that

$$\lim_{n\to\infty} \delta^n(X) = (I - \sharp C^{-1})X.$$

Hence if $C = (\sharp+A)$, the limiting estimator is the optimal linear Bayes estimator. Larger than recommended values of n may, therefore, be useful when accurate information about the tail of the prior is available. For example, if it is thought that the prior has a normal tail, so that $(I-\sharp(\sharp+A)^{-1})X$ is being considered for use, it might pay instead to use δ^n with a large value of n. The resulting estimator will behave similarly to the linear estimator except that it will be more robust with respect to inaccurate prior information. Of course, the larger n is, the less robust δ^n will be.

2. More general classes of priors can be considered. Indeed it can be checked that (2.1) and (2.4) can be replaced by

$$g_n^*(\theta) = \int_0^1 [\det B(\lambda)]^{-1/2} \exp\{-\theta^t B(\lambda)^{-1}\theta/2\} d\mu(\lambda),$$

$$r_n^*(v) = \frac{v \int_0^1 h(\lambda)^{(1+p/2)} \exp\{-vh(\lambda)/2\} d\mu(\lambda)}{\int_0^1 h(\lambda)^{p/2} \exp\{-vh(\lambda)/2\} d\mu(\lambda)},$$

where $B(\lambda) = [C/h(\lambda)] - I$ and $0 < h(\lambda) \leq 1$ for $0 < \lambda \leq 1$. For a wide variety of h and μ, $r_n^*(v)$ can be explicitely evaluated. For example, choosing $h(\lambda) = \lambda$ and $d\mu(\lambda) = I_{(\varepsilon,1)}(\lambda)\lambda^{(n-1-p/2)} d\lambda$ results in a calculable estimator which behaves like $\delta^n(X)$ for small and moderate values of $||X||^2$ (the region depending on ε), but behaves like a linear Bayes estimator for large values of $||X||^2$. As another example, if u is chosen to put unit mass at a particular point, the resulting prior is simply a normal prior. The general class is clearly very rich. (See Efron and Morris (1973a) and Faith (1976) for related classes of estimators in the symmetric situation.) Of the various estimators we considered which arose from priors in this class, δ^n seemed the most attractive. Hence attention was restricted to δ^n.

3. Unfortunately, a problem does arise with δ^* (and with other estimators of the form (1.1)). The estimator definitely performs best when all coordinates are similar or can be transformed so they are similar. (More precisely, this occurs when $[I Q I (I+A)^{-1}]$ is close to a multiple of the identity.) Thus if, for example, there were two groups of similar coordinates, the groups being quite different from each other, it

would probably pay to separately estimate each group. In terms of a prior, this could be interpreted as saying the θ_i's should not be forced to act dependently (as in g_n), but should be separated into independent similar groups, with the resulting prior being say a product of $g_{n_1} g_{n_2}$. The question is - when and how should this separation take place? (Efron and Morris (1973b) give an interesting discussion of the problem in the symmetric situation.)

4. All results in the paper have been for quadratic loss, due to the relative ease of calculation. Numerical studies (such as in Berger (1976b)) have indicated, however, that estimators like δ^* tend to have risks which are quite robust with respect to the functional form (or more precisely the tail) of the loss. See Berger (1976b) for further discussion.

5. The well known relationship between confidence sets and testing of hypothesis, indicates that in some sense the usual multivariate tests of a point null hypothesis can be improved upon by using as an acceptance region $A(\theta) = \{x: \ \theta \in C^*(x)\}$. This question will be pursued elsewhere.

Appendix

Proof of Theorem 3.3.1: The proof is similar to, though considerably more complicated than, the theoretical proof in Brown (1966) of the inadmissibility of the usual confidence sets.

For simplicity, assume that $\Sigma = I$. (This can be assumed without loss of generality as is seen by considering the linearly transformed problem $z = \Sigma^{-1/2}X$, $\eta = \Sigma^{-1/2}\theta$, and $C' = \Sigma^{-1/2}C\Sigma^{-1/2}$.) From Lemma 3.1.2 and the fact that $C \geq \Sigma = I$, it is clear that

$$[1/(n+1)]I \leq \Sigma_n(X) \leq [(2n+1)/(n+1)]I.$$

Hence defining

$$\Omega_\theta = \{x \in R^p: \ \theta \in G^n(x)\} = \{x: \ [\theta-\delta^n(x)]^t \Sigma_n(x)^{-1}[\theta-\delta^n(x)] \leq k(\alpha)\},$$

it is clear that

(A.1) $$x \in \Omega_\theta \Rightarrow |\theta-\delta^n(x)|^2 < k(\alpha)(2n+1)/(n+1).$$

Note next that $\delta^n(x) = x - u(||x||^2)$ (see (3.9) for the definition of u), and that

(A.2) $$u(||x||^2)x \leq 2n|x|/||x||^2 \leq K_1 < \infty.$$

From (A.1) and (A.2) it follows that

(A.3) $$\text{if } x \in \Omega_\theta, \text{ then } |\theta-x| \leq K_2 < \infty,$$

where K_2 can be chosen independent of θ. Assume in the remaining expressions that $x \in \Omega_\theta$, so that (A.3) can be used.

Note first that

$$(A.4)\qquad ||x||^2 = (x-\theta+\theta)^t C^{-1}(x-\theta+\theta) = ||\theta||^2+||x-\theta||^2+2\theta^t C^{-1}(x-\theta)$$

$$= [||\theta||^2+2\theta^t C^{-1}(x-\theta)](1+O(|\theta|^{-1}))$$

$$= ||\theta||^2(1+O(|\theta|^{-1})).$$

From (A.4) it is clear that $||x||^2 > ||\theta||^2/2$ for $|\theta| > K_3$ say, so from Lemma 2.1.1 (xi) it follows that

$$(A.5)\qquad r_n(||x||^2) = 2n + O(|\theta|^{-4}).$$

It can similarly be shown using Lemma 2.1.1 (ix) and Lemma 3.1.1 (iv) that

$$(A.6)\qquad t_n(||x||^2) - r_n^2(||x||^2) = 4n + O(|\theta|^{-4}).$$

From (A.4) it also follows that

$$(A.7)\qquad \frac{1}{||x||^2} = \frac{1}{||\theta||^2} - \frac{2\theta^t C^{-1}(x-\theta)}{||\theta||^4} + O(|\theta|^{-4}).$$

From (A.5), (A.6), and (A.7) it follows that

$$(A.8)\qquad u(||x||^2) = \frac{r_n(||x||^2)}{||x||^2} = \frac{2n}{||\theta||^2} - \frac{4n\theta^t C^{-1}(x-\theta)}{||\theta||^4} + O(|\theta|^{-4}),$$

$$w(||x||^2)C^{-1}xx^tC^{-1} = \frac{(t_n-r_n^2)}{||x||^4} C^{-1}xx^tC^{-1} = \frac{4nC^{-1}\theta\theta^tC^{-1}}{||\theta||^4} + O(|\theta|^{-3}),$$

$$u(||X||^2)C^{-1}x = \frac{2nC^{-1}\theta}{||\theta||^2} + \frac{2nC^{-1}(x-\theta)}{||\theta||^2} - \frac{4n[\theta^tC^{-1}(x-\theta)]C^{-1}\theta}{||\theta||^4} + O(|\theta|^{-3}),$$

and

$$\Sigma_n(X)^{-1/2} = (I - uC^{-1} + wC^{-1}xx^tC^{-1})^{-1/2}$$

$$= (I - \frac{2nC^{-1}}{||\theta||^2} + \frac{4nC^{-1}\theta\theta^tC^{-1}}{||\theta||^4} + O(|\theta|^{-3}))^{-1/2}$$

$$= I + \frac{nC^{-1}}{||\theta||^2} - \frac{2nC^{-1}\theta\theta^tC^{-1}}{||\theta||^4} + O(|\theta|^{-3}).$$

Thus

(A.9) $$\Sigma_n(X)^{-1/2}(\theta-\delta^n(X)) = \Sigma_n(X)^{-1/2}(\theta-x+uC^{-1}x)$$

$$= (\theta-x) + \frac{nC^{-1}(\theta-x)}{||\theta||^2} - \frac{2nC^{-1}\theta\theta^tC^{-1}(\theta-x)}{||\theta||^4}$$

$$+ \frac{2nC^{-1}\theta}{||\theta||^2} + \frac{2nC^{-1}(x-\theta)}{||\theta||^2} - \frac{4n[\theta^tC^{-1}(x-\theta)]C^{-1}\theta}{||\theta||^4} + O(|\theta|^{-3})$$

$$= (\theta-x) + \frac{2nC^{-1}\theta}{||\theta||^2} - \frac{nC^{-1}(\theta-x)}{||\theta||^2} + \frac{2n[\theta^tC^{-1}(\theta-x)]C^{-1}\theta}{||\theta||^4} + O(|\theta|^{-3}).$$

Define

(A.10) $$y = (\theta-x) + \frac{2nC^{-1}\theta}{||\theta||^2} - \frac{nC^{-1}(\theta-x)}{||\theta||^2} + \frac{2n[\theta^tC^{-1}(\theta-x)]C^{-1}\theta}{||\theta||^4}.$$

Clearly (A.10) defines a linear transformation from x to y. The Jacobian of this transformation is

(A.11) $$J = -I + \frac{nC^{-1}}{||\theta||^2} - \frac{2nC^{-1}\theta\theta^tC^{-1}}{||\theta||^2}$$

$$= -(I - \frac{nC^{-1}}{||\theta||^2})(I + \frac{2nC^{-1}\theta\theta^tC^{-1}}{||\theta||^4}) + O(|\theta|^{-4}).$$

As in Lemma 3.2.2, it can be shown that

$$\text{(A.12)} \qquad \det\left(I + \frac{2nC^{-1}\theta\theta^tC^{-1}}{||\theta||^4}\right) = \left(1 + \frac{2n\theta^tC^{-2}\theta}{||\theta||^4}\right).$$

Letting λ_i denote the characteristic roots of C^{-1}, it is also clear that

$$\text{(A.13)} \qquad \det\left(I - \frac{nC^{-1}}{||\theta||^2}\right) = \prod_{i=1}^{p}\left(1 - \frac{n\lambda_i}{||\theta||^2}\right)$$

$$= 1 - \frac{n}{||\theta||^2}\sum_{i=1}^{p}\lambda_i + O(|\theta|^{-4})$$

$$= 1 - \frac{ntr(C^{-1})}{||\theta||^2} + O(|\theta|^{-4}).$$

Combining (A.11), (A.12), and (A.13), it follows that

$$\text{(A.14)} \qquad |\det J|^{-1} = \left[1 - \frac{ntr(C^{-1})}{||\theta||^2} + \frac{2n\theta^tC^{-2}\theta}{||\theta||^4} + O(|\theta|^{-4})\right]^{-1}$$

$$= 1 + \frac{ntr(C^{-1})}{||\theta||^2} + \frac{2n\theta^tC^{-2}\theta}{||\theta||^4} + O(|\theta|^{-4}).$$

Note next from (A.9) and (A.10) that

$$\text{(A.15)} \qquad [\theta-\delta^n(x)]^t\Sigma_n(x)^{-1}[\theta-\delta^n(x)] = |y|^2 + O(|\theta|^{-3})$$

From (A.10) it also follows that $(\theta-x) = y + O(|\theta|^{-1})$, so

$$(\theta-x) = y - \frac{2nC^{-1}\theta}{||\theta||^2} + \frac{nC^{-1}y}{||\theta||^2} - \frac{2n(\theta^tC^{-1}y)C^{-1}\theta}{||\theta||^4} + O(|\theta|^{-3})$$

and

(A.16) $$\exp\{-|\theta-x|^2/2\} = \exp\{-|y|^2/2\}\left[1 + \frac{2ny^tC^{-1}\theta}{||\theta||^2} - \frac{ny^tC^{-1}y}{||\theta||^2}\right.$$

$$\left. - \frac{2n^2\theta^tC^{-2}\theta}{||\theta||^4} + \frac{2n(\theta^tC^{-1}y)^2}{||\theta||^4} + \frac{1}{2}\left(\frac{2ny^tC^{-1}\theta}{||\theta||^2}\right)^2 + O(|\theta|^{-3})\right].$$

Thus from (A.14), (A.15), and (A.16) it can be concluded that

(A.17) $$P_\theta(\theta \in G_n(X)) = \int_{\Omega_\theta} (2\pi)^{-p/2}\exp\{-|x-\theta|^2/2\}dx$$

$$= \int (2\pi)^{-p/2}\exp\{-|x-\theta|^2/2\}|\det J|^{-1}dy$$

$$\{y:\ |y|^2 < k(\alpha)+O(|\theta|^{-3})\}$$

$$= \int (2\pi)^{-p/2}\exp\{-|y|^2/2\}\{1 + \frac{2ny^tC^{-1}\theta}{||\theta||^2} + \frac{ntr(C^{-1})}{||\theta||^2}$$

$$\{|y|^2 < k(\alpha)\}$$

$$\left. - \frac{n||y||^2}{||\theta||^2} - \frac{(2n^2+2n)[\theta^tC^{-2}\theta-(\theta^tC^{-1}y)^2]}{||\theta||^4} + O(|\theta|^{-3})\right\} dy.$$

Defining

(A.18) $$h(\alpha) = (2\pi)^{-p/2} \int_{|y|^2 < k(\alpha)} y_i^2\exp\{-|y|^2/2\}dy,$$

it is easy to check that

$$\int_{|y|^2 < k(\alpha)} \exp\{-|y|^2/2\}(y^tC^{-1}\theta)dy = 0,$$

$$\int_{|y|^2 < k(\alpha)} (2\pi)^{-p/2}\exp\{-|y|^2/2\}||y||^2dy = h(\alpha)tr(C^{-1}),$$

and

$$\int_{|y|^2 < k(\alpha)} (2\pi)^{-p/2}\exp\{-|y|^2/2\}(y^tC^{-1}\theta)^2dy = h(\alpha)\theta^tC^{-2}\theta.$$

It follows from (A.17) that

$$(A.19)\qquad P_\theta(\theta \in C^n(X)) = (1-\alpha) + \frac{n(1-\alpha-h(\alpha))}{||\theta||^2}\{tr(C^{-1}) - \frac{(2n+2)\theta^tC^{-2}\theta}{||\theta||^2}\} + O(|\theta|^{-3}).$$

Defining S_p as the surface area of the unit p sphere, a calculation gives

$$\begin{aligned} ph(\alpha) &= \int_{|y|^2 < k(\alpha)} (2\pi)^{-p/2}|y|^2\exp\{-|y|^2/2\}dy \\ &= (2\pi)^{-p/2}S_p \int_0^{k^{1/2}} r^{(p+1)}\exp\{-r^2/2\}dr \\ &= (2\pi)^{-p/2} S_p[-k^{p/2}\exp\{-k/2\} + p\int_0^{k^{1/2}} r^{(p-1)}\exp\{-r^2/2\}dr] \\ &= -(2\pi)^{-p/2}S_pk^{p/2}\exp\{-k/2\} + p(1-\alpha). \end{aligned}$$

Noting that $S_p = (2\pi)^{p/2}/[2^{(p-2)/2}\Gamma(\frac{p}{2})]$, it follows that

$$(A.20)\qquad h(\alpha) = (1-\alpha) - \frac{k^{p/2}\exp\{-k/2\}}{2^{(p-2)/2}\Gamma(p/2)}.$$

Combining (A.19) and (A.20) gives the desired result except that the error term is $O(|\theta|^{-3})$ instead of $O(|\theta|^{-4})$. In essence, the above argument was a Taylors series argument, and it can be checked that due to the symmetry of the problem the terms which are $O(|\theta|^{-m})$ for m odd must always integrate to zero (as did the term $[2ny^tC^{-1}\theta/||\theta||^2]$). Hence the next nonzero term of the

expansion for $P_\theta(\theta \in G^n(X))$ will be $O(|\theta|^{-4})$.

Proof of Theorem 3.3.4: A very laborious calculation exactly paralleling the proof of Theorem 3.3.1 (but including all terms up to $O(|\theta|^{-4})$) gives in place of (A.19)

$$(A.21) \qquad P_\theta(\theta \in G^n(X)) = (1-\alpha) + \frac{(p-2)}{8(\theta^t \Sigma^{-1} \theta)^2} \{4p(p-2)\{1-\alpha) - 2(p-2)(3p+4)h(\alpha)$$

$$+ [p^3+3p^2-30p+16]\ell(\alpha)/3 - (p-1)[p^2+2p-32]g(\alpha)\},$$

where $h(\alpha)$ is defined in (A.18), and

$$\ell(\alpha) = (2\pi)^{-p/2} \int_{|y|^2 < k(\alpha)} y_1^4 \exp\{-|y|^2/2\} dy,$$

$$g(\alpha) = (2\pi)^{-p/2} \int_{|y|^2 < k(\alpha)} y_1^2 y_2^2 \exp\{-|y|^2/2\} dy.$$

These integrals can be explicitely evaluated ($h(\alpha)$ is evaluated in (A.20)), giving

$$\ell(\alpha) = 3(1-\alpha) - \frac{6[k/2]^{p/2}\exp\{-k/2\}}{p\,\Gamma(p/2)} \left(1 + \frac{k}{2(p+2)}\right),$$

$$g(\alpha) = (1-\alpha) - \frac{2[k/2]^{p/2}\exp\{-k/2\}}{p\,\Gamma(p/2)} \left(1 + \frac{k}{(p-1)}\left[1 - \frac{3}{2(p+2)}\right]\right).$$

Inserting these expressions in (A.21) and collecting terms gives the desired result. ||

[1] Berger, J. (1976a). Admissible minimax estimtion of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist. 4, 223-226.

[2] Berger, J. (1976b). Tail minimaxity in location vector problems and its applications. Ann. Statist. 4, 33-50.

[3] Berger, J. (1976c). Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. J. Multivariate Anal. 6, 256-264.

[4] Berger, J., and Bock, M. E. (1977). Improved minimax estimators of normal mean vectors for certain types of covariance matrices. S. S. Gupta and D. S. Moore (Eds.) Statistical Decision Theory and Related Topics II, Academic Press, 1977.

[5] Berger, J., and Srinivasan, C. (1977). Generalized Bayes estimators in multivariate problems. To appear in the Ann. Statist.

[6] Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. Ann. Math. Statist. 37, 1087-1136.

[7] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist. 42, 855-904.

[8] Brown, L. D. (1974). An heuristic method for determining admissibility of estimators - with applications. Technical Report, Rutgers University.

[9] Casella, G. (1977). Minimax ridge regression estimation, Ph.D. thesis, Purdue University, May 1977.

[10] Dempster, A. P., Schatzoff, M., and Wermuth, N. (1976). A simulation study of alternatives to ordinary least squares (with discussion). J. Amer. Statist. Assoc. 72, 77-106.

[11] Efron, B. and Morris, C. (1973a). Stein's estimation rule and its competitors - an empirical Bayes approach, J. Amer. Statist. Assoc. 68, 117-130.

[12] Efron, B. and Morris, C. (1973b). Combining possibly related estimation problems. J. Roy. Statist. Soc., B, 35, 379-421.

[13] Efron, B. and Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. 4, 11-21.

[14] Faith, R. E. (1976). Minimax Bayes set and point estimators of a multivariate normal mean. Technical Report No. 66, University of Michigan.

[15] Goldstein, M. (1976). Bayesian analysis of regression problems. Biometrika 63, 51-58.

[16] Haff, L. R. (1976). Minimax estimators of the multinormal mean: autoregressive priors. J. Multivariate Anal. 6, 265-280.

[17] Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55-68.

[18] Hoerl, A. E., Kennard, R. W., and Baldwin, K. R. (1975). Ridge regression: some simulations. Commun. Statist. 4, 105-123.

[19] Hudson, M. (1974). Empirical Bayes estimation. Technical Report #58, Stanford University.

[20] Joshi, V. M. (1967). Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. Ann. Math. Statist. 38, 1868-1875.

[21] Judge, G. and Bock. M. E. (1977). Implications of pre-test and Stein rule estimators in econometrics. North Holland Publishing Company, the Series Studies in Mathematical and Managerial Economics.

[22] Leonard, T. (1976). Some alternative approaches to multiparameter estimation. Biometrika 63, 69-75.

[23] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. J. Roy. Statist. Soc. B, 34, 1-41.

[24] Morris, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series. Rand Corporation, California.

[25] Olshen, A. (1977). Comment on "A note on a reformation of the S-Method of multiple comparison" by H. Scheffé. J. Amer. Statist. Assoc. 72, 144-146.

[26] Rao, C. R. (1977). Simultaneous estimation of parameters - a compound decision problem. S. S. Gupta and D. S. Moore (Eds.) Statistical Decision Theory and Related Topics II, Academic Press.

[27] Rolph, J. E. (1976). Choosing shrinkage estimators for regression problems. Commun. Statist. A5(9), 789-802.

[28] Rubin, H. (1977). Robust Bayesian estimation. S. S. Gupta and D. S. Moore (Eds.) Statistical Decision Theory and Related Topics II, Academic Press, 1977.

[29] Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Prob. 1, 197-206. University of California Press.

[30] Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution. J. Roy. Statist. Soc. B, 24, 265-296.

[31] Stein, C. (1973). Estimation of a mean of a multivariate distribution. Proc. Prague Symp. Asymptotic Statist. 345-381.

[32] Stein, C. (1974). Estimation of the parameters of a multivariate normal distribution - I. Estimation of the means. Stanford University, Department of Statistics, Technical Report No. 63.

[33] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. Ann. Math. Statist. 42, 385-388.

[34] Strawderman, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances. Ann. Statist. 1, 1189-1194.

[35] Thisted, R. A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Ph.D. Thesis, Stanford University.

[36] Zellner, A. and Vandaele, W. (1971). Bayes-Stein estimators for K-means, regression and simultaneous equation models. Studies in Bayesian Econometrics and Statistics (S. Feinberg and A. Zellner, Eds.) Amsterdam: North-Holland Publishing Co.